A study of diachronic changes in the productivity of several Early New High German derivational morphemes using the RIDGES corpus

Masterarbeit
zur Erlangung des akademischen Grades
Master of Arts (M.A.)
im Fach Linguistik

Humboldt-Universität zu Berlin
Sprach- und literaturwissenschaftliche Fakultät
Institut für deutsche Sprache und Linguistik

Eingereicht von	Pankratz, Elizabeth
Wissenschaftliche Betreuerin	Prof. Dr. Anke Lüdeling
Ort und Datum	Berlin, den 16.07.2019

Acknowledgements

I could not have done this project alone, so I want to say thank you to those who helped guide me in writing this thesis: Anke Lüdeling, for your supervision, for suggesting a topic and then giving me the freedom to take it and run. Felix Golcher, for help with the analyses. Jackson Berry, Timo Buchholz, Jan Fliessbach, Philip Loewen, Allen Pankratz, David Pankratz, and the audience of the HU's corpus linguistics colloquium, for valuable discussions and suggestions. And Raúl Bendezú Araujo, for the pep talks and all the afternoons spent writing over coffee and lemon cake.

Thank you also to those who have played an important role in my last two years: Uli Reich, Bob van Tiel, and Stefan Müller, for seeing something in me. Roland Schäfer, for teaching me a great deal of what I know. Franziska Sattler, for your companionship, your inspiration, and for being my biggest cheerleader. And my loved ones back home, for your support and presence in spite of the distance.

Finally, thanks are also due to the Ausländerförderung of the Konrad-Adenauer-Stiftung, without whose financial support and network of international students these two years would have been much more difficult.

Contents

1 Introduction

Language users have a strong, intuitive sense about the acceptability of derived words in their native language. An English native speaker will accept the derivation of *serene* with *-ity* to create *serenity*, as well as the derivation of *calm* with *-ness* to create *calmness*, but probably will not accept *calm* with *-ity*: **calmity*.

Even though both of these suffixes, *-ness* and *-ity*, are used to derive nouns from adjectives, this example shows that they cannot be applied equally freely to all adjectives; further constraints must exist. As Aronoff (1976: 35) says: "Though many things are possible in morphology, some are more possible than others." This difference in the possible ways that derivational morphemes can be used is a phenomenon that linguists often refer to as morphological productivity.

Roughly put, productivity has to do with the potential of a derivational morphological process to create a new word (cf. Bauer 2001: 10, C. Scherer 2007: 260, Dal & Namer 2016: 70). Intuitively, the process *-ness* seems like it has a greater potential to be used for new words than *-ity*. Can this intuition about the relative productivity of these two morphological processes also be quantified?

Finding a way to pin down morphological productivity enough to put a number on it has been the subject of much research in the last decades, primarily by R. Harald Baayen and his colleagues. An oft-heard comment in the literature is that productivity is a complex, multifaceted phenomenon, and it is best measured using a combination of several different statistics (cf. e.g. Plag 1999: 30, C. Scherer 2007: 259, Hartmann 2016: 146). The accumulated inventory of productivity measures has been summarised excellently in Zeldes (2012: Chapter 3).

In addition to comparing the productivity of derivational processes in modern language, we may also want to look at the way in which their productivity has changed over time. This adds another dimension of complexity to the problem. Most measures focus on quantifying productivity as a "potential or a value which is attached to a word-formation rule at a particular point in time" (Cowie & Dalton-Puffer 2002: 417), i.e. they analyse productivity synchronically. Applying these synchronic measures to discover something about diachronic change in productivity generally involves dividing the historical period of interest into subperiods, calculating the measures on the subcorpus for each subperiod, and then comparing the results, concluding from a change in the value produced by the measure that the process' productivity changed over time (cf. Cowie & Dalton-Puffer 2002: 421; this method is used by Doerfert 1994, Cowie 1999, C. Scherer 2005, Schneider-Wiejowski 2011, Hartmann 2016, and Kempf 2016, among others).

This method crucially presupposes that the measures be sensibly comparable between the different subcorpora that represent each subperiod. More often than not, though, this is not the case. Subcorpora tend to be differently sized, since not every historical period gave rise to the same amount of text, and it is unusual for corpora to contain the same amount of text for every subperiod. This is a problem for the usual method described above, because the main measures used in diachronic productivity studies – type counts, hapax legomena, vocabulary growth curves, potential productivity, and the parameter S from the finite Zipf-Mandelbrot LNRE model – cannot actually be sensibly compared between differently-sized subcorpora. The primary goal of this thesis is to illustrate why not and to propose an alternative method for future work on changing productivity.

To set the stage, I will discuss in Section 2 the ways in which the literature approaches morphological productivity, and I will also illustrate why we might want to study it diachronically using the example of the role of morphology in register changes in Early New High German (ENHG). Section 3 will explain how I gathered the ENHG data used for the Section 4 illustration of the diachronic productivity measures mentioned above. After exploring the problematic behaviour of these measures, I will suggest in Section 5 that other, more complex but also more cognitively plausible, statistical methods for modelling productivity probabilistically offer a more fruitful way forward.

2 Why study diachronic productivity?

This section will motivate the topic of this thesis, first zooming out to give an overall view of how the literature approaches the study of productivity and how the current quantitative approach fits in. Then, I will consider a specific research question about the role that morphology plays in the development of a new scientific register in ENHG, as an example of the sort of question that successful measures of diachronic productivity would help us explore.

2.1 Approaches to productivity

We can conceive of a derivational morphological process as a sort of function that accepts a suitable base lexeme as input and produces as output, i.e. derives, a different lexeme that contains the derivational morpheme of interest (Spencer 2016: 27, Booij 2012: 53). This conceptualisation is schematised in Figure 1.

This schema is helpful because it lets us differentiate various ways of looking at productivity. To discover something about a morphological process' productivity, which is understood as a characteristic of the process itself (Plag 1999: 2), we can analyse

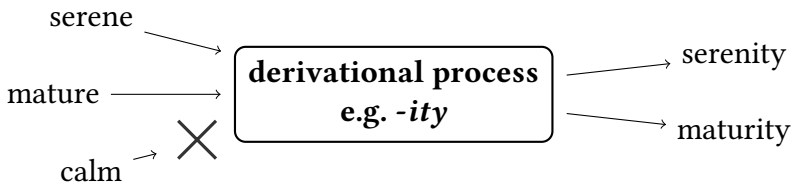

Figure 1: An example schema of a derivational process that accepts some (but not all) bases as input and produces derived words as output.

the process' input (the left side of the schema) and/or the output (the right side of the schema).

On the one hand, we may ask what is possible as input and consider the restrictions that prevent, for instance, *calm* from being selected as suitable input to the process *-ity*. This research is mostly descriptive and qualitative, outlining the phonological, morphological, lexical, syntactic, and semantic restrictions that an input must fulfill in order to be derived with the process at hand (Cowie & Dalton-Puffer 2002: 414, Bauer 2001: 128–137). Many of the studies that look at morphological productivity in historical corpora focus on this side of the schema, describing qualitatively what serves as input to a morphological process and how this changes over time (e.g. Demske 2000, Doerfert 1994, Wegera & Prell 2000).

On the other hand, we can also analyse a process' output to learn about the way the morphological process is actually used (Baayen 2001, 2003, 2009). This approach is quantitative, and it is on this side of the schema that the productivity measures to be discussed in this thesis are situated.

Both sides of a morphological process – the input and the output – are essential pieces of the puzzle (cf. Lüdeling, Evert & Heid 2000: 57, Cowie & Dalton-Puffer 2002: 413), and it is one of the primary drawbacks of the measures to be discussed below that they only consider the process' output in trying to quantify its productivity. I will return to this point in the final discussion. Before I move on to discuss the sampling process and the measures themselves, a brief sketch of ENHG and its derivational morphology is in order.

2.2 Derivational morphology in Early New High German

ENHG was spoken between about 1350 and 1650 in regions which are today part of Germany, Denmark, Poland, Austria, Italy, Switzerland, and France (W. Scherer 1878: 13, Hartweg & Wegera 2005: 30; different estimates for ENHG's periodisation are conveniently charted in Hartweg & Wegera 2005: Figure 3). During this time, a gradual linguistic transition was taking place in the sciences: academics and professionals were

shifting from Latin to German as the primary language of science (cf. Klein 2010). In the fourteenth and fifteenth centuries, German had widely been considered "barbaric" and was relegated to domains outside of schools and universities, but over the following centuries, it developed into an important medium of academic discussion, finally reaching widespread acceptance in the sciences toward the end of the seventeenth century (Klein 2010: 498).

This development is interesting with respect to derivational morphology, because the use of derivational morphemes played an important role in the expansion of ENHG into the sciences. For ENHG to become functional in its new scientific register, new German words had to be formed to express concepts that were previously only lexicalised in Latin (cf. Klein 2010, Habermann 2001). Speakers filled these gaps by applying derivational processes to the existing words, expanding ENHG's vocabulary to meet their emerging communicative needs (Klein 2010: 491, Baayen et al. 2011: 462). Having quantitative tools at hand to explore historical changes in productivity would allow us to explore the changing morphology of this period more deeply.

The use of derived words in ENHG – especially words derived to become nouns, i.e. nominalisations – was also promoted by Latin style rules that were adopted into German (Habermann 2001). For example, a noun phrase headed by an abstract nominalisation was preferred over a subordinate clause, because the condensed information was considered particularly elegant: "*Wenn man sagt: ich wundere mich, daß du so unbeständig bist … heißt es eleganter: ich bestaune deine Unbeständigkeit*" 'If you say: I admire how capricious you are … this can be said more elegantly as: I marvel at your capriciousness' (Rössing-Hager 1990: 411, cited in Habermann 2001: 24, my translation).

The role of derivational morphology in ENHG is therefore a changing one over the centuries, apparently playing a more prominent role as time goes on (cf. Ruh 1956, Brendel et al. 1997, Habermann 2001). We should be able to find productive uses of derivational morphology by examining texts from this period, which would allow us to study changing productivity in action.

The texts most likely to contain instances of productively formed nominalisations are botanical or herbal texts. These were around in German as early as the fifteenth century (Klein 2010: 499-500), in contrast to many other texts that were also written on scientific topics. The reason for botanicals' early appearance in German was that they contained medicinal information that was useful to the general (German-speaking) public, so they were made available in German far earlier than many other genres of scientific text. What this means for us today is that botanical texts were being written in ENHG during the entire period of the development of ENHG's new scientific register, and so we expect to see the scientific register development reflected there (cf. Odebrecht

et al. 2017). Looking at the evolution of morphology in botanical texts can shed light on the general case: how ENHG morphology changed to accommodate the demands of the developing register. With this goal in mind, 61 botanical texts from between 1482 and 1914 were collected in the RIDGES Herbology corpus (Lüdeling, Odebrecht, et al. n.d., Odebrecht et al. 2017). The data used in this thesis – samples of three ENHG nominalising morphemes, *-er*, *-heit*/*-keit*, and *-ung* – was gathered from this corpus.

Ideally, we would like to discover whether there is an increase of productivity in nominalising processes during the ENHG period, which would reflect the increasing need to coin new words in order to fill the gaps in ENHG's vocabulary. Another reason to expect an increase in nominalisations is because they tend to be more frequent in educated registers, since they express a lot of information very succinctly (Plag, Dalton-Puffer & Baayen 1999: 212). Also, Wegera & Prell (2000: 1594) note a general increase across ENHG in using derivation rather than a syntagma to express complex ideas. We would need reliable measures of productivity in order to confirm this hypothesis, though, and as we will see, these are not at hand.

However, we do know quite a bit about the behaviour of ENHG nominalising morphemes through descriptive work done by Wegera & Prell (2000), Doerfert (1994), C. Scherer (2005), Demske (2000), Müller (1993), and others. For the rest of this section, I will briefly summarise the known qualitative characteristics of the three suffixes under investigation. For each one, I will give some examples from RIDGES and impressionistic notes about how it is used there. (A detailed qualitative analysis of these suffixes in RIDGES exceeds the scope of the present project, but this should certainly be done in future.)

2.2.1 *-er*

In ENHG, the suffix *-er* was used mainly to form agent and instrumental nouns (Wegera & Prell 2000: 1595). C. Scherer (2005: 46) adds that it could also create toponymic nouns and words for objects and abstract states of affairs. Verbs, nouns, and toponyms (i.e. place names) may belong to its set of inputs, to create nominalisations such as the following (sampled from RIDGES): *betteler* 'beggar' from the verb *betteln* 'to beg', *gärtner* 'gardener' from the noun *Garten* 'garden', and *Passawer* 'person from Passau' from the toponym Passau. In RIDGES, many of the *-er* derivations are words for plants, e.g. the flowering plant *Schmetterlingsblümler* (lit. 'butterfly bloomer').

It has the extended forms *-ner* and *-ler*, as in *Afrikaner* 'person from Africa' from *Afrika* 'Africa', or *Künstler* 'artist' from *Kunst* 'art'. Sometimes it also induces an umlaut on the stem, like in *Künstler* or *Träger* 'carrier' from *tragen* 'to carry' (cf. C. Scherer 2005).

C. Scherer (2005: 144) finds that in the *Mainzer Zeitungskorpus*, verbs are most common part of speech for the base, followed by nouns and then toponyms, while Müller (1993: 237) shows that in texts by Albrecht Dürer, verbal and nominal bases are approximately equally frequent.

2.2.2 *-heit/-keit*

The pair of suffixes *-heit*/*-keit* is presented together, following Doerfert (1994), because they are two allomorphs of the same morpheme. (Another allomorph *-igkeit* exists too, but this one is left out of the label for clarity's sake.) The primary function of this morpheme in ENHG was to derive words for people's physical and mental characteristics and states of being, as well as for behaviours and actions (Doerfert 1994: 296–297, Wegera & Prell 2000: 1595). It mainly took adjectives as input – e.g. *Ewigkeit* 'eternity', from *ewig* 'eternal' – which were gradually joined by past participles – e.g. *Vergessenheit* 'obscurity' from *vergessen* 'forgotten' (cf. Wegera & Prell 2000: 1595).

Beyond *-heit*, *-keit*, and *-igkeit*, there used to be other variants including *-cheit*, *-icheit*, and *-ikeit*, though these all disappeared before the first half of the sixteenth century (Wegera & Prell 2000: 1595, Doerfert 1994: 291). None of these older variants appear in RIDGES, but the diphthongs do occur in other spellings; see Section 3.1 below.

2.2.3 *-ung*

Finally, *-ung* was a nominaliser of primarily verbal bases to produce nouns for events, resulting states, and objects (Hartmann 2016: 90, Wegera & Prell 2000: 1596, Demske 2000: 367). Examples of these from RIDGES include *sterckung* 'strengthening' from *stärken* 'to strengthen', *Wachsung* 'growth' from *wachsen* 'to grow', and *wonung* 'living space' from *wohnen* 'to live'.

According to Demske (2000: 368–369), *-ung* appeared preferentially on morphologically complex bases (prefixed and particle verbs) over simplexes and transitive verbs over intransitive ones. This seems largely in harmony with the results from RIDGES, though some simplex derivations that are impossible for modern German also appear in the sample, such as *Machung* 'making' from *machen* 'to make'. The unacceptability of *machen* as input today is a clear illustration that the behaviour of the derivational process *-ung* has changed over time, and with a deeper dive into the samples I have gathered, one could surely find out a great deal about the qualitative changes in each morpheme's productivity.

dipl	ein	reines	Glåſzlin	erflieſſen	.	Rectifici-	rung	vnd	lew⸗	terũg	der	waſ⸗
clean	ein	reines	Gläszlin	erfliessen	.	Rectificirung		vnd	lewterumg|lewterung		der	was-
norm	ein	reines	Gläslein	erfließen	.	**Rektifizierung**		und	Läuterung		der	Wasser

Figure 2: A screenshot of the text tiers *dipl*, *clean*, and *norm* in RIDGES (link to this match in ANNIS: https://tinyurl.com/rektifizierung).

3 Gathering the samples from RIDGES

As mentioned above, the corpus I am using for this study of productivity measures is the RIDGES Herbology corpus in Version 8.0 (Lüdeling, Odebrecht, et al. n.d.), which contains 61 botanical texts written between 1482 and 1914, comprising 257,537 tokens (Belz et al. 2018). RIDGES stands for "Register in Diachronic German Science", and it was created for the purpose of exploring how the ENHG scientific register developed over time (cf. Odebrecht et al. 2017).

In more accurate terms, 257,537 is the number of tokens found on the *dipl* annotation tier. RIDGES supplements the original texts with many annotation levels that contain a great deal of information, from part of speech data to the type of lettering used in the original document (cf. Odebrecht et al. 2017). The three levels of annotation I used for my study are the text tiers *dipl*, *clean*, and *norm*.

The tier *dipl* is a diplomatic transcription of the text that preserves textual elements from the original, e.g. line breaks, mid-word hyphenations, and archaic graphemes such as the long S <ſ> and umlauted vowels like <å>, as in *Glåfzlin* 'little glass'.

Two levels of normalisation follow *dipl*. First, *clean* tidies up the historical spellings, for instance changing *Glåfzlin* into *Gläszlin*, and combining words that were divided across line breaks into a single unit. Both of these levels are more or less direct reflections of the textual information in the original texts; the first step involving some interpretation comes in the next tier, *norm*. Here, the various historical spellings are mapped to forms written according to modern German orthographic standards, so that *Gläszlin* becomes *Gläslein*. The graphematic normalisation allows the user to search for, e.g., the word *Kräutern* 'herb' (dative plural) on *norm* and obtain all instances of this word in any of its variations, like *Kråutern*, *Kreutern*, *Kreuttern*, or *Kråuteren* (Odebrecht et al. 2017: 704). Figure 2 shows these three text tiers as displayed in the ANNIS corpus architecture through which RIDGES can be accessed (Krause & Zeldes 2016).

I used ANNIS to query RIDGES and export the results as .csv files, which I then manually tidied and lemmatised; the rest of this section will explain the procedures I used for these three steps. The queries themselves as well as details about export parameters are included in Appendix A.

Suffix	Graphematic variants
-heit/-keit	*-heyt/-keyt*, *-hait/-kait*
-ung	*-ug*, *-nng*, *-umg*, *-unn*

Table 1: The graphematic variants for each suffix appearing in RIDGES.

3.1 Querying

To collect as many tokens derived with *-er*, *-heit/-keit*, or *-ung* as possible, I had to consider different graphematic and inflectional forms that might appear in the corpus and structure my queries accordingly. This section explains how I went about this.

The sample for *-er* was assembled from queries on *clean* for word-final <er> and word-final <ern> (the dative plural), since there was no graphematic variation to take into account (cf. C. Scherer 2005: 88). These separate queries were merged while tidying.

The suffixes *-heit/-keit* and *-ung* do show graphematic variation, in the diphthong and the nasal respectively. I looked first for the graphematic variants that appear in RIDGES by taking an overall sample of all of the words ending in *-heit/-keit* or *-ung* on *norm*, and then looking at the *dipl* and *clean* tiers to see which variations on the normalised spelling might appear in the original texts.

Because of the way this query had to be structured to retrieve both *dipl* and *clean*, it may have excluded certain tokens that were divided into two units on *dipl* but only one on *clean* (which would be the case if e.g. the token were split over two lines in the original document; see the division of *Läuterung* on *dipl* in Figure 2). For this reason, to create the basis sample for the analysis, I did a second sample of all of the *-heit/-keit* and *-ung* derivations on *norm* with only *clean* in addition, leaving out *dipl*. Even though *clean* would be the basis for lemmatisation, I searched on *norm* because of the unified orthography.

Then, I did separate queries to see if any tokens containing the graphematic variants that I found in the first exploratory sample were not picked up in the sample for analysis (i.e. did not end with *-heit*(*en*), *-keit*(*en*), or *-ung*(*en*) on *norm*). None of these cases turned up additional tokens that the first sample had missed, so we can be sure that for words that originally contained either of these suffixes, the graphematic normalisation reflects the original morphology accurately. (The other way around is a different story, since some words not originally derived with these suffixes were normalised to contain them; I will come to this in the next section.)

The graphematic variation found in RIDGES for each of these suffixes is shown in Table 1.

On *clean*	On *norm*
runde	*Rundung* ‘rounded form/area’
erfarnüssz	*Erfahrung* ‘experience’
vergifft	*Vergiftung* ‘poisoning’

Table 2: Examples of forms normalised to contain a derivational suffix that the original does not have.

3.2 Tidying

The queries for all three suffixes also returned some false positives: tokens containing strings with the same form as the suffixes that are not actually instances of this morphological process. In the tidying stage I manually removed these from each sample.

The forms *-heit*/*-keit* had the fewest false matches, since they are quite distinctive, while *-er* had the most false matches, because many German morphemes share this same form, including a plural morpheme and several adjectival suffixes. The search for word-final <er> also returned many proper names, which were removed based on context.

The sample for *-ung* also needed some tidying to remove false matches like *jung* ‘young’ and *Ursprung* ‘source’. Interestingly, several now-defunct nominalisations on *clean* were normalised as containing *-ung*, because *-ung* is still viable today, in contrast to the processes in these older words. Examples are listed in Table 2. These were also removed from the sample.

3.3 Lemmatising

At this point, the data returned by my query was trimmed down to tokens that contain the derivational process I am interested in, but the tokens also contain other morphological processes in addition to the one I want to study. For example, the derivations might appear in their plural forms, as heads of compounds, and so on, and leaving these other processes in the sample would distort the statistics. Consider that counting a bare derivation and a derivation in the head of a compound as two different types would lead to artificially higher counts for unique words, i.e. types, and for words appearing only once in the sample, i.e. hapax legomena. For this reason, the tokens appearing on *clean* for each sample were lemmatised in a way that reflects only the derivational processes being used, following Lüdeling, Evert & Heid (2000) and Hartmann (2016).

It is not always easy to decide what is at the base of a word containing many morphological processes and thus what should count as its own type (cf. Hartmann 2016: 159, Plag 1999: 28, C. Scherer 2005: 40). I drew guidance from Lüdeling, Evert & Heid (2000: 59–60) and Hartmann (2016: 160–165), and the guidelines I used are as follows:

1. Various (mis)spellings of the same word, such as *Fäulung* and *Faulung*, are all counted as the same type, here *Faulung* 'decay, putrefaction'.
2. Forms prefixed with *un-* are generally considered as the type of the version without *un-*, e.g. *Ungesundheit* 'non-health' is lemmatised as *Gesundheit* 'health'.
3. Compounds are considered as the same type as their head, e.g. the tokens *Gattung* 'species', *Pflanzengattung* 'plant species', and *Pilzengattung* 'mushroom species' all belong to the type *Gattung*, and *Lohnkutscher* 'driver of a rental carriage' is an instance of the type *Kutscher* 'driver of a carriage'.
4. Prefixed forms that were already prefixed before derivation, e.g. particle verbs, are considered as their own types. For example, *Herstellung* is not the same type as *Stellung*, because *Herstellung* 'manufacturing' cannot be analysed as *her* 'from' + *Stellung* 'position', nor is *Zusammensetzung* 'composition' the same type as *Setzung* 'setting', since the two are semantically quite different.[1]
5. Prefixed forms that are likely instances of prefixation after derivation, e.g. *Zusammenvermischung* 'mixture together', are considered as the type of the non-prefixed version, here *Vermischung* 'mixture'.

The tokens from *clean* that were lemmatised in this way form the basis of all the analyses to follow.

4 Methods for quantifying diachronic productivity

This section will explore several of the most frequently-used measures used to quantify diachronic productivity. I will discuss their mathematical behaviour to show why they fail to be useful, comparable measures across differently-sized subcorpora, and explain where their deeper conceptual problems lie.

The discussion that follows is relevant for any studies that involve comparing morphological productivity, whether diachronically between subperiods or synchronically between different morphological processes (or a combination of both).

The data I will use to test the behaviour of these measures are the samples for *-er*, *-heit*/*-keit*, and *-ung* from RIDGES. For each suffix, Table 3 presents the number of tokens (i.e. all words containing each morpheme), types (i.e. all unique words containing each morpheme), and hapax legomena (i.e. those words that only appear once).

[1] Hartmann (2016) notes that it should also be considered whether the bare derivation, e.g. *Setzung*, ever appears on its own. If not, it does not make sense from a usage-based perspective to consider *Setzung* as a type. This pertains more to larger corpora, though, since in small corpora like RIDGES, the absence of *Setzung* in the sample does not mean it was never used in the language. Semantic criteria are therefore more viable.

Suffix	Tokens	Types	Hapaxes
-er	270	105	65
-heit/*-keit*	652	126	57
-ung	1992	391	191

Table 3: Type, token, and hapax legomenon counts from RIDGES for *-er*, *-heit*/*-keit*, and *-ung*.

The analyses that follow were conducted in R (R Core Team 2014), with the help of the `zipfR` package developed by Stefan Evert and Marco Baroni (Evert & Baroni 2007, Baroni & Evert 2014).[2]

4.1 Type counts

One way to look at how productive a morphological process has been in the past is to consider how many different words contain that process up until whenever the measurement was taken (cf. Bauer 2001: 48–49, Cowie & Dalton-Puffer 2002: 416, Zeldes 2012: 94). The logic behind this is that a productive process is likely to have created more unique words, i.e. types, than an unproductive process, so by comparing the type count between different processes or subperiods, we should be able to say something about the relative productivity of each process in each time period.

A note about terminology: In the literature and in what follows, type count is sometimes referred to as V for "vocabulary", and token count will be represented with the variable N or referred to as the sample size.

Simply plotting the number of types for each suffix over time creates the graph in Figure 3. Naively analysing this graph would lead one to conclude that around 1750, *-ung* starts appearing in many more types than previously, so it must start to become more productive around that time, while the other two suffixes show no such change.

However, drawing conclusions from type counts in this blind fashion is misleading, because type counts are a sample-based measure, and the number of types must always be understood in relation to the number of tokens that contain them (cf. Zeldes 2012: 55). A greater number of tokens is likely to contain a greater number of types, so that an increase in type count (like the one we see for *-ung* after about 1750) might not truly be indicative of an increase in use, but rather an artefact of differently-sized samples or subcorpora (where a subcorpus might consist of a single text, or several texts that all come from a certain subperiod).

[2]The samples and scripts used for my analyses will be accessible on GitHub under the following URL: `github.com/epankratz/diachronic-productivity-thesis`.

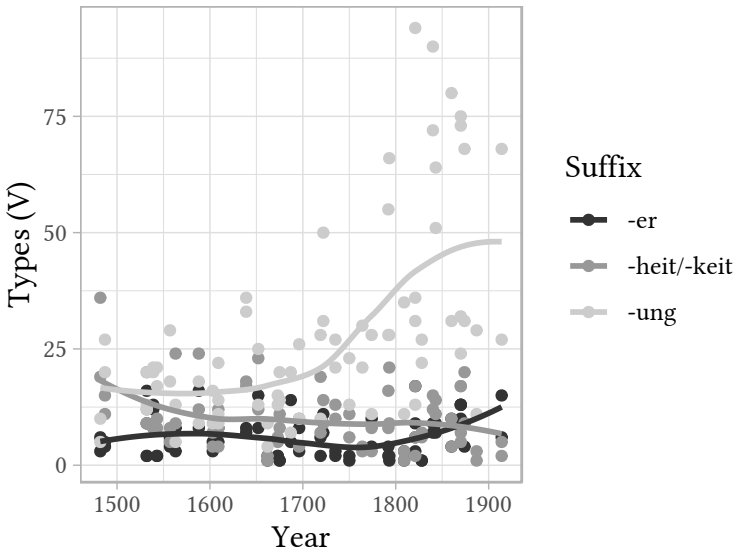

Figure 3: The number of types for each suffix in RIDGES by year. Each data point represents one text.

A smaller subcorpus will contain fewer tokens overall, which will likely be made up of fewer, mostly high-frequency types, in contrast to larger subcorpora, which have the opportunity to pick up more low-frequency types due to the larger sample size.

From this, it follows that observed type counts cannot sensibly be compared between differently-sized subcorpora. One way forward is simply to make an educated judgment by considering the length of the texts that the types come from.

The graph in Figure 4 shows the development of text length in RIDGES over time.[3] The trend in text length is to decrease as time goes on. This is noteworthy in conjunction with Figure 3, since it indicates that the increased type count of *-ung* from about 1750 onward is not simply due to text length effects. The reason for its sudden increase in usage is probably that, starting around the same time, RIDGES begins to include medicinal texts, which likely use this suffix to a different extent than the more traditional botanical texts.

A slightly more sophisticated measure than overall type counts per subcorpus would be to consider the number of *new* types that appear in each successive subcorpus. The idea is that, among the types appearing in one subcorpus that we have not seen in any of the previous subcorpora, some might have been productively formed during the corresponding subperiod (Cowie & Dalton-Puffer 2002: 431). With this measure, we might be able to get a sense of how many new words were coined by the morphological process of interest during this time (cf. Bauer 2001, Cowie 1999, Cowie & Dalton-Puffer 2002).

[3]Sometimes a single text is divided into several separate documents in RIDGES; all documents from one text are combined into a single data point for these analyses.

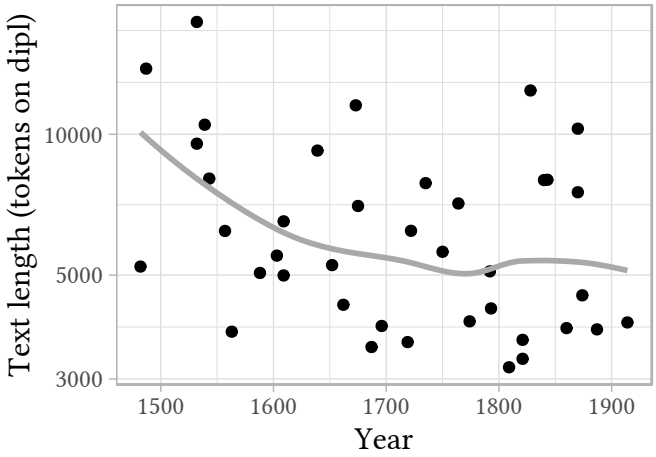

Figure 4: Text length in terms of number of tokens on the *dipl* annotation tier for each text in RIDGES by year.

If there were many new types – which are assumed to reflect the number of newly formed words – then we might conclude that the process is being applied more productively in that time period. On the other hand, if only the same old words that already existed were being used and no new words showed up, then one could conclude that the morphological process was not being productively used. Considered diachronically, an increase in the number of new types between two subperiods is understood as an increase in productivity, and the closer to zero the rate of addition gets (because no new types appeared in the given subcorpus), the lower the productivity.

The first subcorpus in which one measures the rate of addition will be characterised by an infinite increase in the number of new types, since there is no previous subperiod to compare to. This is avoided by taking all the types appearing in the first (or first few) subperiods and creating from these an initial lexicon as a basis for comparison for the next subperiod. The lexicon then accumulates the new types as they are encountered in each subperiod/subcorpus. This method is used by Cowie & Dalton-Puffer (2002: 430–431) and Cowie (1999: 118).

The new type counts of all three suffixes over 100-year subperiods is shown in Fig 5. The first subperiod would be equivalent to the starting lexicon, since all types encountered there are new.

This method is still problematic, though, since there is no way to know whether a newly encountered word is completely new to the language, whether it is well-known and lexicalised and just happens to show up here for the first time, or whether it is somewhere in between (cf. Cowie 1999: 76–77). So, new type counts must be understood only as an approximation, but the approximation is rather worse for smaller corpora, since the smaller samples have less opportunity to pick up words, meaning that more

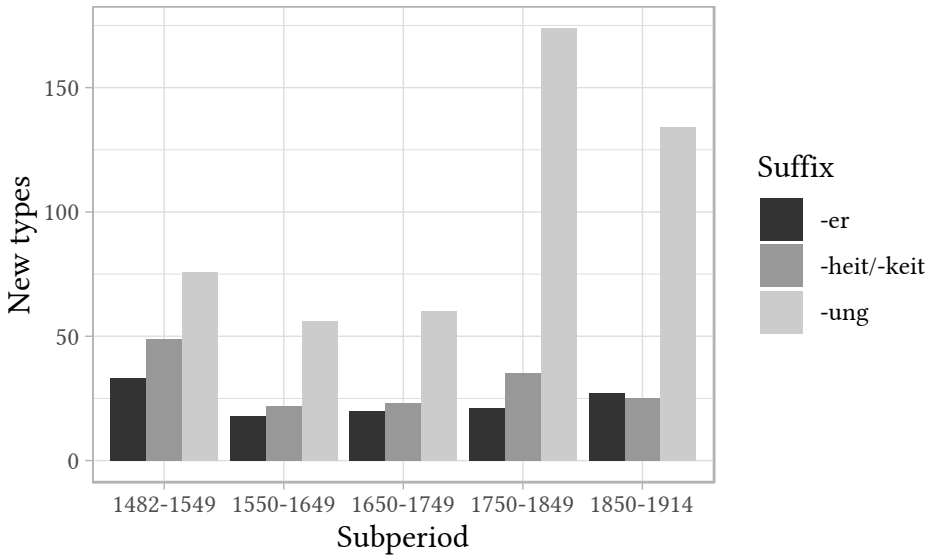

Figure 5: New type counts per subperiod for *-er*, *-heit*/*-keit*, and *-ung*.

known words are likely to show up as "new" between subperiods. And since larger subcorpora have a greater statistical likelihood to pick up more new types, an increase in new types could just as well be due to an increase in sample size as to an actual diachronic effect, and this still cannot be satisfactorily teased apart.

A strategy proposed to remedy this and to rescale type counts so that they are comparable between differently-sized subcorpora is normalisation, which will be the focus of the next section.

4.1.1 Normalisation

Normalisation is usually done to avoid the problem of comparability across differently sized subcorpora (cf. Cowie & Dalton-Puffer 2002: 427). Consider Table 4, which shows the number of tokens on *dipl* and the number of types containing *-ung* in the earliest five texts in RIDGES. Based only on this data, it is difficult to discern whether *-ung* is being used more or less between these documents (and consequently, whether the productivity can be understood as increasing or decreasing). There is a sudden upswing in type count between the first document and the second, but this might just be reflective of the second text being much larger. This means that the actual type counts cannot be compared without somehow adjusting them so that they can all be understood on the same scale.

The tool for adjusting type counts in this way is normalisation. Normalisation involves decreasing the type count's value if the types come from a large subcorpus, and augmenting the value if the types come from a smaller subcorpus, such that all absolute

Title	K_a	V_a
Das Buch der Natur	5,215	5
Gart der Gesundheit	13,817	27
Artzney Buchlein der kreutter	9,550	12
Contrafayt kreüterbuch	17,387	20
New Kreütter Buch	10,484	21

Table 4: Actual corpus size in *dipl* tokens K_a and actual type counts V_a for *-ung* derivations in the first five texts in RIDGES.

type counts are converted "into *n* per 1000 or 10 000 words" (Cowie & Dalton-Puffer 2002: 427).

In what follows, I will explore the very simple mathematics of the normalisation procedure used by e.g. C. Scherer (2005) in her study of *-er* and Hartmann (2016) in his study of *-ung*, so that we can more easily see the problem with it.

For each subcorpus (which, in a per-document analysis, is equivalent to a text), we first take the ratio of the actual type count, V_a, to the actual corpus size (here, text length) it comes from, K_a. We equate this ratio to another one with the normalised corpus size of our choosing, K_n (normally a pleasantly round number like 10,000, which is what I use here), in the denominator, and the unknown value of the normalised type size, V_n, in the numerator, as shown in Equation 1 below. In other words, we want to find the number V_n which has the same ratio to K_n as V_a has to K_a. By setting the two ratios equal to one another and then solving for V_n, we get 2, which is the formula used for calculating normalised token counts given the other three constants.

$$\frac{V_n}{K_n} = \frac{V_a}{K_a} \tag{1}$$

$$V_n = \frac{V_a \times K_n}{K_a} \tag{2}$$

Superficially, this seems to make it possible to compare normalised type counts V_n out of K_n tokens overall, since now all counts are scaled in the same way, namely as V_n types out of $K_n = 10{,}000$ tokens. For illustration, after normalising the data from Table 4 in this way, we see in Table 5 that the jump from the first text to the second is now only a twofold leap rather than a fivefold one. So, it is still an increase, but one that is not as dramatic as we might have assumed if we had not rescaled the values to compensate for original corpus size.

Title	K_a	V_a	V_n
Das Buch der Natur	5,215	5	10
Gart der Gesundheit	13,817	27	20
Artzney Buchlein der kreutter	9,550	12	13
Contrafayt kreüterbuch	17,387	20	12
New Kreütter Buch	10,484	21	20

Table 5: Actual corpus size K_a, type counts V_a, and normalised type counts scaled to $K_n = 10{,}000$ V_n, for *-ung* derivations in the first five texts in RIDGES.

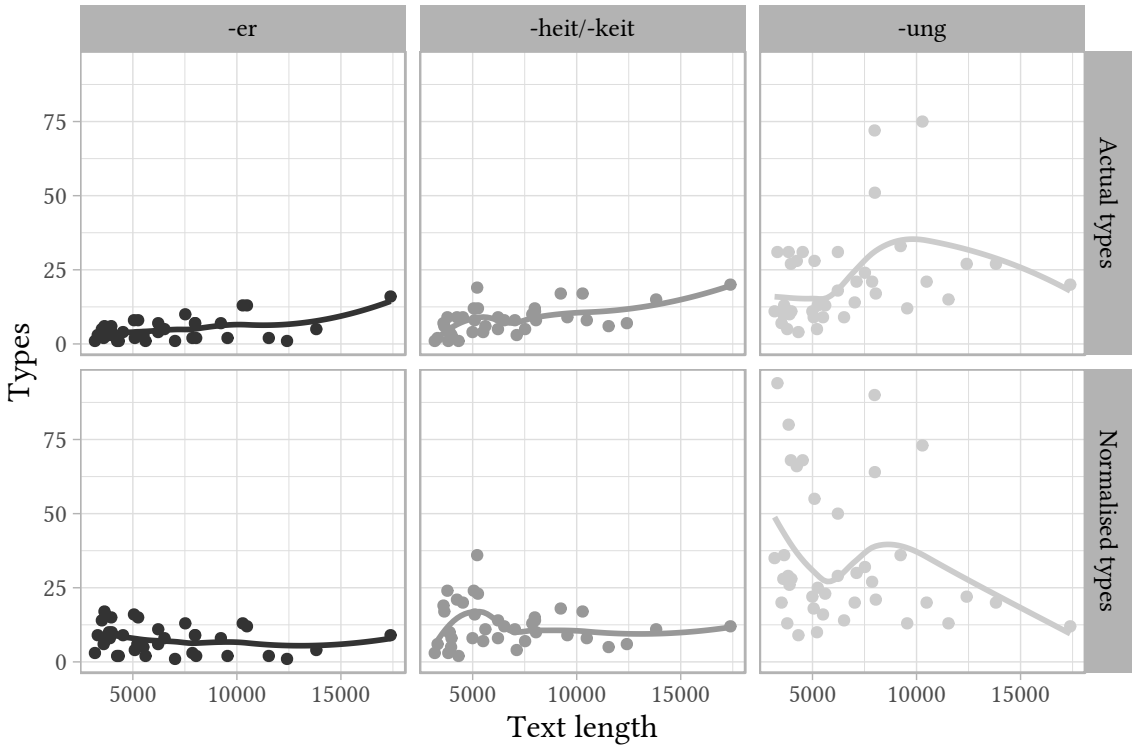

Figure 6: Actual and normalised type counts by original corpus size for each suffix. Each data point represents one text.

The plot in Figure 6 shows the effect of normalisation on the samples for *-er*, *-heit*/*-keit*, and *-ung*. The data is presented with text length on the x axis to make especially apparent the effect that normalisation has of increasing the value of the type counts originating from shorter texts and decreasing the value of type counts originating from longer texts. This is particularly clear for *-er* and *-heit*/*-keit*, where the smoothing line has a positive slope in the upper row (for the non-normalised type counts) but is more or less horizontal in the lower row (for the normalised type counts).

The reason that *-ung*'s distribution looks so unlike the other two is because it occurs especially frequently in texts from 1750 onward, and those texts also tend to be shorter, hence the clustering in the lower left of its xy plane.

Even though the type counts now look to be scaled comparably, effects from the original sample are still at play. Type counts normalised to K_n tokens are still dependent on their original corpus size K_a, which means that we should actually not be comparing normalised type counts from differently-sized subcorpora either (cf. Lüdeling 2009: 337).

This makes sense when we consider how the normalisation expression in Equation 2 above works. The original corpus size in the denominator of this expression is what equalises the expression's value. A large denominator (i.e. a large original corpus size) gives the fraction a relatively smaller value than a smaller denominator (i.e. a small original corpus size). This inflates the normalised type count for types coming from smaller subcorpora and deflates the normalised type count for types from larger subcorpora.

However, for this same reason – the original corpus size being in the denominator – the normalised type count is dependent on the size of the original subcorpus that it comes from. The larger the original subcorpus, the relatively smaller the resulting type count will be. Graphically, we see this dependency as a tendency of the curve toward zero in Figure 7 (cf. Lüdeling 2009: 337, Figure 1), and mathematically, it can be described as

$$\lim_{K_a \to \infty} \frac{V_a \times K_n}{K_a} = 0.$$

This dependency is problematic for historical corpus studies, since the number of total tokens in a corpus K_a is unlikely to be the same between subcorpora from different subperiods, as mentioned above. Hartmann (2016: 168) justifies the use of normalised counts in his study of *-ung* by stating that he is more interested in identifying general trends over the entire period contained in the corpus, rather than making precise comparisons between two neighbouring subperiods. For comparable K_as this may still be reasonable, but when K_as vary widely, normalisation should be done with care.

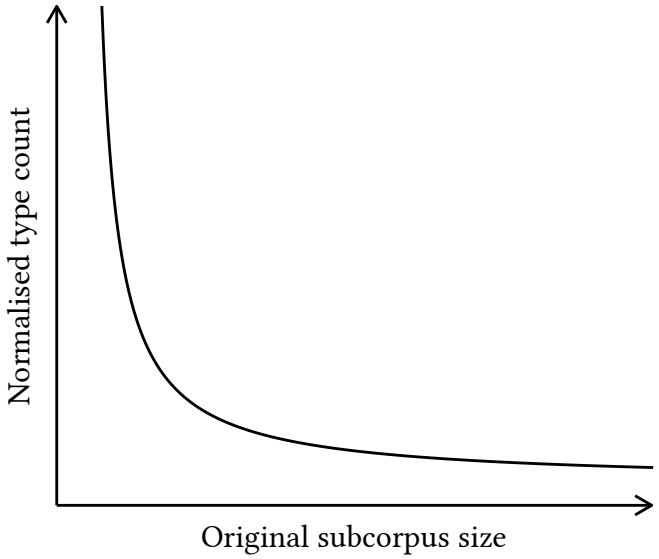

Figure 7: An idealised plot of how the normalised type count V_n decreases as K_a, the actual original corpus size it is sourced from, increases.

Also, as we saw above, larger subcorpora are likely to contain more tokens containing the morphological process of interest, and thus possibly more types. Cowie & Dalton-Puffer (2002: 427) suggest that the normalised counts for the subcorpora that were originally large are still likely to be larger than the counts for the subcorpora that were originally smaller, so normalisation does not account for the likelihood of a greater type count from larger (sub)corpora. However, the dependency on original corpus size shown in Figure 7 might actually compensate for this by decreasing the number of types from larger corpora. The question of to what extent this dependency might actually counteract large corpus effects should be explored in future in a more principled way.

The next section moves on from exploring the past productivity of a suffix in its use in corpus data to looking at its potential to be used in new coinages, which is another key part of the behaviour of a morpheme that we would like to describe (and one that, as we will see, is difficult to quantify).

4.2 Hapax legomena and potential productivity

How can corpus data be used to quantify the potential that a morphological process had at any given time to create new words? This section will explore this question by introducing hapax legomena as an operationalisation of newly coined words and the measure potential productivity from Baayen (2001, 2009), referred to as $\mathscr{P}$, as a probability that the next token we encounter is a newly (and thus productively) formed word.

If we want to look at the word-formation behaviour of a morpheme, the words that we are most interested in are neologisms, those freshly-minted words that have just been derived using that morphological process. These were not specifically considered by the type count measure above because it treats all types the same, whether they appeared once or a hundred times. Newly-formed words are more telling about a morpheme's productivity than the total number of types with that morpheme, because if many new words are coined in a given period, it follows that the morphological process that coined them was probably quite productive in that period. We would therefore like to find out how many neologisms there are in each subperiod and compare this number over time.

We approximate the number of words that are likely to be neologisms using the number of hapax legomena, i.e. types that only appear once in the sample (cf. Zeldes 2012: 60–61). Hapaxes are accepted as an approximation because "productively created items are one-off unique occurrences, and therefore they must form a subset of the hapax legomena" (Zeldes 2012: 60). Also, hapaxes could be cognitively relevant; if a speaker hears many unique words that contain a particular morpheme, they might conclude that they themselves can also use that morpheme to create unique words. On the other hand, if they never hear any unique words containing a morpheme, only the same old words over and over, they might learn that they cannot use that morpheme in innovative ways.

Hapax legomena are only an imperfect approximation of neologisms, though. Not all neologisms appear as hapaxes: A language user might productively form a word in order to talk about a particular concept at hand, and the reference using this new word might happen several times, and as soon as the new word is used again, it is no longer detectable by counting words that only occur once (Zeldes 2012: 60–61). Also, not all hapax legomena will correspond to neologisms, since already-coined, well-known words might happen to only show up once in the sample.

In short, the set of hapax legomena and the set of productively formed words are likely to overlap, but unlikely to coincide completely. Hapaxes should therefore only be understood as a heuristic for measuring productivity, rather than a direct measure of it (cf. Dal & Namer 2016: 74, Zeldes 2012: 60–61).

Also, in small corpora, the overlap of the set of neologisms and the set of hapaxes is even more tenuous. As Cowie (1999: 78) says, "the smaller the corpus, the more types will be hapax legomena", meaning that the set of hapax legomena in a small corpus will probably contain many known words which happen to appear only once (cf. also Cowie & Dalton-Puffer 2002: 416, Hartmann 2016: 166). Also, since the samples are smaller, they are less likely to pick up the actual neologisms that were really used in that subperiod (of which we have no idea how many existed). For these reasons, Cowie & Dalton-Puffer (2002: 431) and Cowie (1999: 116) argue against using hapaxes in small

corpus studies to gauge changes in productivity. In fact, Cowie (1999) leaves them out of her analysis completely, working only with new type counts by period.

If we accept the assumption that hapaxes map fairly well to neologisms, though, we can use the number of hapaxes in certain other measures to tell us something about the propensity of a morphological process to show up in further neologisms. A measure that is widely used to do this, potential productivity, will be the focus of the next section.

4.2.1 Potential productivity as a probability

The measure called potential productivity $\mathscr{P}$ is formulated as follows (cf. Baayen 2001: 50), where "in category C" means the derivations containing a particular morpheme.[4]

$$\mathscr{P} = \frac{\text{number of hapax legomena in category } C}{\text{number of tokens in category } C} = \frac{\text{hapaxes}}{N} \tag{3}$$

This ratio is intuitive if we consider it as a probability. The proportion of hapax legomena encountered in a sample to the number of tokens in that sample seems like a reasonable estimate of the probability that the next token encountered will also be a hapax legomenon (Baroni 2009: 818). If we accept the operational stand-in of hapaxes for neologisms, this will tell us the likelihood that the next token will be a neologism and thus the likelihood with which a morphological process will produce new members (cf. Zeldes 2012: 63).

The interpretation of $\mathscr{P}$ is very simple: its value ranges between zero and one. It is zero when the numerator is zero, i.e. when there are no hapax legomena among all of the tokens derived with a morphological process, and this is interpreted as the process being completely unproductive in that sample, because there are no new formations using it. On the other hand, $\mathscr{P}$ evaluates to one when every token in the category of interest is a hapax legomenon, so the numbers in the numerator and denominator are the same; the process would then be considered perfectly productive, since every word is freshly coined. In reality, the values sit most often between these two extremes, but the rule of thumb is that the closer the value is to one, the more productive the process is at that point in the sample.

The caveat "at that point in the sample" is crucial, though, because a process' "probability of producing previously unseen forms … decreases the more we have seen of a process's output" (Zeldes 2012: 64). In other words, the value of $\mathscr{P}$ decreases as N, the number of tokens in category C, increases.

[4] Other names for this measure include "category-conditioned degree of productivity" or "productivity in the narrow sense", and it is sometimes represented simply as P, e.g. in Baayen (2009), probably for typographical reasons (cf. Zeldes 2012: 63).

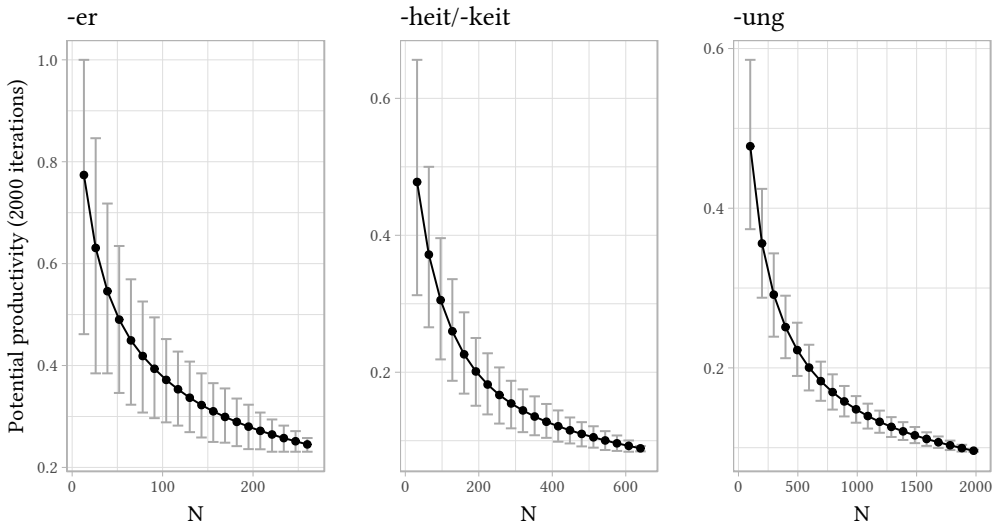

Figure 8: Monte Carlo means and confidence intervals after 2,000 interations for *-er*, *-heit*/*-keit*, and *-ung*.

This dependency on token count has the same pattern as the dependency observed above for normalised data on original corpus size. In both cases, the ratio's value gets smaller as N increases. The changing denominator N means that the resulting values are actually scaled differently, so they cannot be directly compared across samples with different Ns.

I will illustrate this dependency of $\mathscr{P}$ on N for the data I gathered from RIDGES by calculating Monte Carlo means and confidence intervals following the method outlined in Tweedie & Baayen (1998).

For each of the three suffixes *-er*, *-heit*/*-keit*, and *-ung*, I first randomised the order of all of the tokens in the sample (in order to obscure any possible actual diachronic effects, since I am interested here in the theoretical, mathematical behaviour of $\mathscr{P}$; cf. Tweedie & Baayen 1998, Hartmann 2016). Then, at twenty equally-spaced measurement points throughout the sample, each of which added $\frac{N}{20}$ tokens to the sample size, I calculated $\mathscr{P}$ for the entire sample up to that point.

For example, for *-ung* where the total $N = 1{,}992$, each measurement point was ninety-nine tokens apart, so the first value for $\mathscr{P}$ was calculated at $N = 99$, the second at $N = 198$, and so on.

This randomisation and calculation procedure was done 2,000 times to create a distribution of 2,000 $\mathscr{P}$ values for each measurement point. The resulting means and the 95% Monte Carlo confidence intervals (which are simply the innermost 95% of values at each measurement point, with the top and bottom 2.5% removed) are plotted for each suffix in Figure 8.

These plots illustrate clearly the dependency of $\mathscr{P}$ on N, namely that $\mathscr{P}$ decreases as N increases. If the value of $\mathscr{P}$ were conceivably constant over all values of N, we would expect that a horizontal line would be compatible with the confidence range across increasing N, but this is not the case.

(The reason that the error bars get smaller and smaller as N increases is because of the increasing homogeneity of the samples as N gets closer and closer to the actual sample size. To see why, consider that there will be much more variability in the sample at the first measurement point. Randomly taking 99 tokens from *-ung*'s sample of 1,992 tokens 2,000 times will produce 2,000 very different samples, while randomly taking 1,881 tokens from 1,992 will result in 2,000 largely similar samples.)

Hartmann (2016: 166–167) does a similar investigation, creating 336 subcorpora out of the 336 texts in the GerManC corpus (also in a randomised order, so that no actual diachronic effects play into his illustration), where each subsequent subcorpus contains one more text than the one before it. He calculates and plots $\mathscr{P}$ for each growing subcorpus, also showing that $\mathscr{P}$ approaches zero as N increases.

This characteristic of $\mathscr{P}$ is important to have in mind for diachronic studies of productivity, because dealing with different sample sizes is the norm when looking at productivity over time. Since $\mathscr{P}$ scales differently for different values of N, we cannot compare $\mathscr{P}$ between differently-sized subcorpora.

Consider what would happen with two hypothetical subcorpora, the first with 300 tokens and the second with 1000 tokens. A higher value for $\mathscr{P}$ in the first subperiod compared to the second might be due to a larger number of hapax legomena in the first subperiod than in the second (from which we would conclude that the process' productivity was higher in the first period than in the second). However, this situation could also arise if the number of hapax legomena were exactly the same, since the value for $\mathscr{P}$ gets smaller as the denominator gets larger, regardless of the constant in the numerator.

A constant number of tokens in each subperiod would be necessary in order to accurately compare $\mathscr{P}$ in a diachronic corpus (cf. Hartmann 2016: 168). Some methods for achieving this will be discussed in Section 4.5 below. But first, to further illustrate the behaviour of $\mathscr{P}$ and to set the stage for the final sort of measure that aims to quantify the potential of a process to coin new words – LNRE models – in Section 4.4, the following section will introduce one more tool for our toolkit: the vocabulary growth curve.

4.3 Vocabulary growth curves and potential productivity

A vocabulary growth curve or VGC is a useful tool for exploring the make-up of a sample in terms of types and tokens. It tracks the way that the type-token ratio changes as one

moves token by token through a sample (to understand the name, recall that the type count is sometimes represented by V for "vocabulary"). This is helpful because the VGC looks different for productive and unproductive morphological processes, which I will illustrate below.

We are interested in the type-token ratio, shown in Equation 4, because it can tell us something about how salient the morphological process is. From a smaller type-token ratio, i.e. few types and many tokens, we might presume that the types that exist are lexicalised and thus that the word formation pattern might be recognisable, but neither particularly salient nor productive. On the other hand, a larger type-token ratio could indicate a greater productivity, since there are more distinct types compared to all the tokens (cf. Hartmann 2016: 146).

$$\text{type-token ratio} = \frac{\text{types in } C}{\text{tokens in } C} \tag{4}$$

By considering the number of types and tokens in an entire sample, the type-token ratio looks at the "end result" of that sample. In contrast, a VGC looks at how the sample arrives at this end result. VGCs have the advantage of being able to describe a sample without being dependent on the size of that sample.

As an illustration, consider Figure 9, which shows the N and V values for the first twenty *-er* derivations in RIDGES and the corresponding VGC, which emerges when N is plotted along the x axis and V along the y axis.

Theoretically, the shape of a VGC created in this way can be used to indicate the productivity of a morphological process. Let us imagine two scenarios, one where we are tracking the vocabulary growth of a prototypically productive process, and one where we are tracking the growth of a prototypically unproductive process. A productive derivational process can be consistently used to form new words, which would each count toward the number of new types we encounter as we proceed through the sample. So, as the sample size gets bigger, we will keep encountering more and more new words that we have not seen before, because they keep being productively formed, which means that the VGC will continue to grow.

On the other hand, if the process is not productive, no new types will be formed using it, so the sample will eventually reach a point where the entire vocabulary of types containing that process has already been seen. No matter how large the sample gets, every new instance of the category is just a reuse of one of the types we have already seen, because no new types can be formed. In this situation, the slope will become zero and the curve will flatten out. These two scenarios would produce VGCs resembling those in Figure 10.

Token	*N*	*V*
Zauberer ‘wizard’	1	1
Schreiber ‘writer’	2	2
Zauberer ‘wizard’	3	2
Sünder ‘sinner’	4	3
Schöpfer ‘creator’	5	4
Schöpfer ‘creator’	6	4
Schöpfer ‘creator’	7	4
Maler ‘painter’	8	5
Schöpfer ‘creator’	9	5
Zeiger ‘pointer’	10	6
Lehrer ‘teacher’	11	7
Gärtner ‘gardener’	12	8
Frierer ‘freezer/cold’	13	9
Zeiger ‘pointer’	14	9
Persier ‘Persian’	15	10
Bettler ‘beggar’	16	11
Lehrer ‘teacher’	17	11
Weber ‘weaver’	18	12
Spanier ‘Spaniard’	19	13
Apotheker ‘pharmacist’	20	14

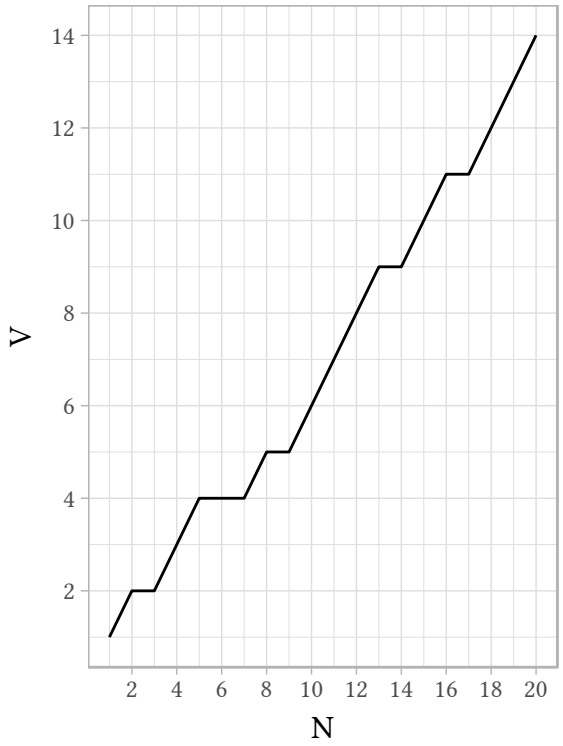

Figure 9: Type and token counts and a vocabulary growth curve for the first twenty *-er* derivations in RIDGES.

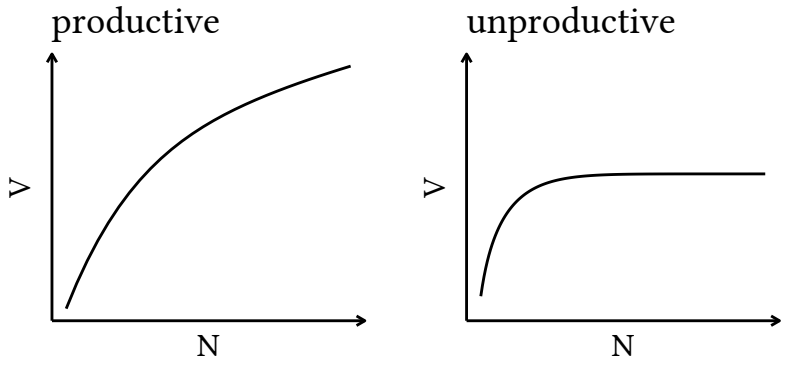

Figure 10: Idealised vocabulary growth curves representing typical productive and unproductive processes.

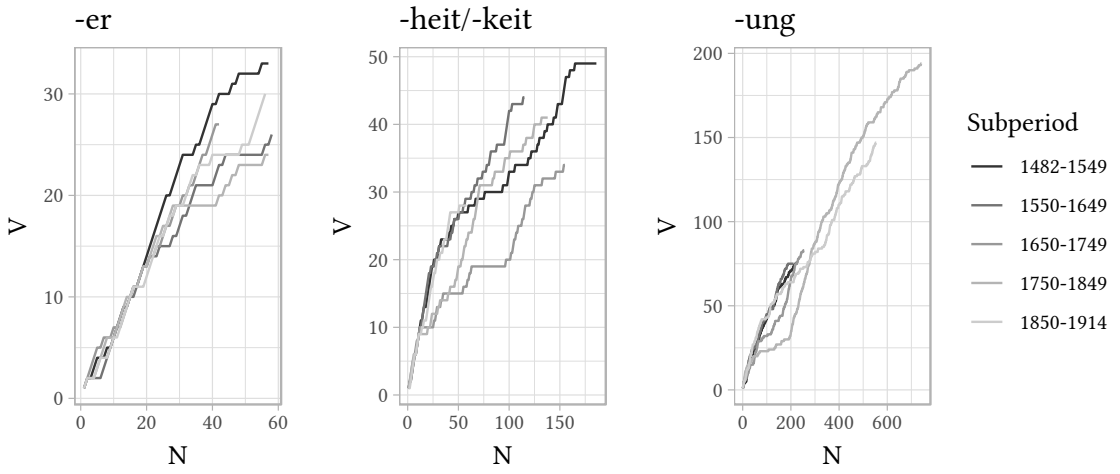

Figure 11: Observed vocabulary growth curves for each suffix in 100-year subperiods.

The advantage of VGCs, as mentioned above, is that they can describe a sample independent of the sample size. However, their drawback is that the curves all end once all tokens in the sample have been exhausted, and unless a curve has already begun to obviously flatten out (indicating that we have stopped encountering new types), we cannot really conclude anything about the productivity of the processes based only on the VGC itself. This is especially problematic for small samples, since their VGCs only show steep initial increases, which could just as well be the beginning of a productive or an unproductive curve. We might be only a few types away from exhausting the full vocabulary that this category has to offer, or the curve might continue upward and consistently encounter new types; we simply do not know (cf. Evert & Lüdeling 2001).

Comparing characteristics of the sections of the VGCs that we do have is also unfruitful. A steeper slope does not indicate higher productivity, but only that new types are occurring in quicker succession for one process than for another, which is more an indication of the process' frequency than of its productivity (see Bauer 2001: Section 3.4 for further discussion of the frequency/productivity distinction).

In reality, VGCs of token-by-token steps through the actual sample practically never proceed along an ideal curve. The empirical VGCs for 100-year subperiods for *-er*, *-heit*/*-keit*, and *-ung* in RIDGES are shown in Figure 11, and they are not as smooth as the idealised curves presented in Figure 10, since they contain all the artefacts of words' non-random distributions in texts (cf. Baroni & Evert 2014: 7). To obtain a smoother curve (which can then be manipulated using techniques from calculus), Baayen (2001: Chapter 2) uses binomial interpolation to create so-called "interpolated VGCs". This procedure is implemented for R by Evert & Baroni (2007) in the `zipfR` package.

An interpolated VGC gives the expected values of the vocabulary size, i.e. the expected type count $E[V]$, up to any value of N which is less than or equal to the orig-

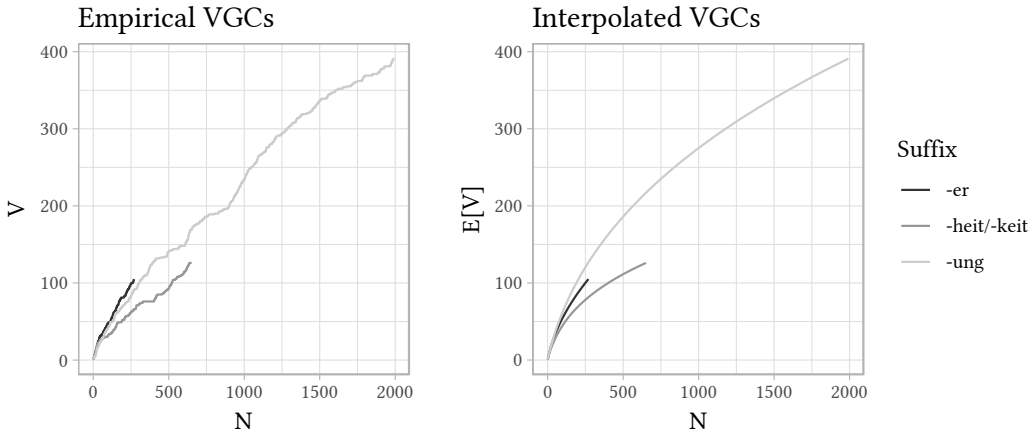

Figure 12: Empirical and interpolated vocabulary growth curves for the full samples of *-er*, *-heit*/*-keit*, and *-ung*.

inal sample size. Baroni & Evert (2014: 7) explain that "[t]hese expected values can be thought of as the average of vocabulary size … computed over a large number of randomizations of the order of tokens in the corpus." The empirical and interpolated vocabulary growth curves for the entire samples of *-er*, *-heit*/*-keit*, and *-ung* are shown for illustration in Figure 12.

In actual fact, the binomial interpolation procedure seems to achieve the equivalent of randomising the order of tokens by simply disregarding it. By taking as input only a list of how many types correspond to how many tokens in the sample (a so-called "frequency spectrum", illustrated for *-er* in Table 6), the observed rate of vocabulary growth and actual distribution of types in the sample are obscured.

The assumption that language can be appropriately modelled as randomised words is one that can certainly be disputed, and I will return to this point in the next section.

For now, though, we can use this new tool of empirical and interpolated VGCs to discuss the mathematics behind the potential productivity measure introduced above.

4.3.1 Potential productivity as a slope

In Section 4.2.1, I introduced potential productivity $\mathscr{P}$ as the number of hapax legomena of category C divided by the number of tokens in C (see Equation 3) and discussed how it can be understood as a probability. This is true, and it provides an intuitive way to understand the ratio, but now that we have seen how VGCs work, we can explore $\mathscr{P}$'s mathematical nature as well. This will give us a deeper understanding of why comparing $\mathscr{P}$ across samples with different numbers of tokens is ill-advised.

Baayen (2001), who first introduced $\mathscr{P}$, shows that it is the slope of the tangent line to the interpolated VGC at the VGC's endpoint, i.e. at the maximum value of N for

m	V_m
1	54
2	16
3	6
4	8
5	1
7	1
9	1
10	1
13	1
14	1
15	1
19	1

Table 6: The frequency spectrum for the full sample of *-er* in RIDGES, where m represents the token frequency ($m = 1$ are hapax legomena, $m = 2$ are dis legomena, etc.) and V_m represents the number of types that have that same frequency, i.e. the sample contains 54 hapax legomena, 16 dis legomena, and so on, up to the most frequent type that occurs 19 times (cf. Baroni 2009).

that sample (see Baayen 2001: 50–51 for more mathematical detail). In plainer terms, $\mathscr{P}$ is a measure of how quickly the VGC is growing after all tokens have been sampled (Baayen 2009: 902). This last bit is crucial, because the point at which all tokens have been sampled will be different between differently-sized samples. This means that the measurement of $\mathscr{P}$ will happen at different points in different VGCs, and changing how far the VGC extends along the x axis will change the value for $\mathscr{P}$. Its slope depends on where the endpoint is located (Evert & Lüdeling 2001: 2), and it will get shallower and shallower as we proceed farther through the sample (Zeldes 2012: 64). Framed in this way, we see why $\mathscr{P}$, as the slope of the VGC at its endpoint, should not be compared for different curves.

Furthermore, knowing the particular value of $\mathscr{P}$ at the endpoint of the VGC of interest actually tells us very little about the curve itself. The slope begins as one and approaches zero and must pass through every value in between one and zero as the curve flattens out (by the intermediate value theorem for continuous functions; Renze & Weisstein n.d.). Thus, since every value of the slope $\mathscr{P}$ is contained in every curve, the resulting value of $\mathscr{P}$ is not informative about the productivity of the process overall, especially since conceivably, two different curves might have the same slope, i.e. the same value for $\mathscr{P}$, at the same value of N. For illustration, consider Figure 13. $\mathscr{P}$ can only tell us the rate of accumulating new types at the specific point where all tokens in the sample have been seen, but it is not informative beyond this point in any way.

This limitation of interpolated VGCs and $\mathscr{P}$ is theoretically overcome by the LNRE

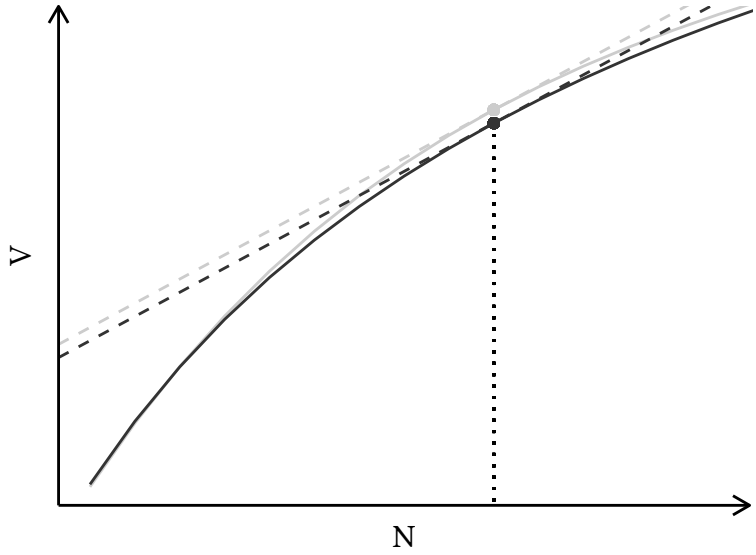

Figure 13: A schematic illustration of two different vocabulary growth curves that have the same slope (i.e. the same value of $\mathscr{P}$) at the same value of N.

models put forward in Baayen (2001), which have the ability to extrapolate the VGC from the observed frequency spectrum to any arbitrary value for N. The motivation for this is that, once we have VGCs that extend arbitrarily far along the x axis, we should be able to figure out how the productivity of several processes compares by comparing these curves. This idea and its implementation will be the focus of the next section.

4.4 The finite Zipf-Mandelbrot LNRE model

Recall the ideal productive and unproductive VGCs presented in Figure 10. The curve for the productive process looks like it is consistently continuing to grow (albeit increasingly slowly), in contrast to the curve for the unproductive process, which quickly approaches some horizontal asymptote. Even though we cannot see it in these plots, we can reasonably assume that the productive process will also start "flattening out" at some point, approaching a horizontal asymptote situated higher on the y axis than that of the less productive process (cf. Evert 2004).

It might initially seem unintuitive that a productive process will have a maximum number of types that it could create. After all, we understand a high productivity to mean that a morphological process can reliably take in base words to create new output. And as new words enter the language over time, some of these will likely be suitable as input for a process, so that their derivations become new types created by that process. This would mean that it does not make sense to set a limit for the number of types a productive process can create, because new types are consistently being produced.

If we would keep expanding our sample over time, adding new words as they enter the language, then it would indeed be inaccurate to assume such a maximum number of types. However, this is not how we normally analyse language diachronically. Common practice in historical corpus linguistics is to consider all of the language use (that we can access) in a particular time as one "chunk" that can be compared against other chunks (Cowie & Dalton-Puffer 2002). And these chunks of language are finite, because only so much language could have been produced during each period. Within a finite chunk of language, the set of available inputs for even the most productive process will also be finite, so the process will eventually derive all of the existing inputs, producing from these some maximum number of types. This is why it makes sense to assume a finite potential vocabulary even for highly productive processes.

One might object by suggesting that output from other productive morphological processes could serve as input to the process of interest, which could potentially make this input infinite. For example, the modern German *-igkeit* appears on words already derived using *-los*, e.g. *Sprachlosigkeit* 'speechlessness' and *Aussichtslosigkeit* 'hopelessness', so the output set of the "subordinate" process *-los* feeds the input set of the "superordinate" process *-igkeit*.

However, this does not lead to an infinite input set for the superordinate process, because the finiteness argumentation applies transitively. A subordinate process whose output is acceptable input for a superordinate process would also only have a finite set of inputs, and once these are exhausted, the subordinate process will have produced all possible outputs that it can. This finite set of outputs then augments the finite set of inputs for the superordinate process, rather than expanding it infinitely.

Could compounding infinitely extend the input set of a process, since there are no real restrictions on compound formation? At least for the question of derivational morphology, I would argue no, since we analyse compounds only as the type of their heads rather than as a new type each time (see Section 3.3 above and Lüdeling, Evert & Heid 2000). For this reason, compounds whose head is of a particular type will not actually add to the input set, since this set consists of types, and each compound is an individual token of the type that is the compound's head (e.g. *Gattung* 'species', *Pflanzengattung* 'plant species', and *Pilzengattung* 'mushroom species' are three tokens of the type *Gattung* and therefore do not separately increase the size of the input set of *-ung*).

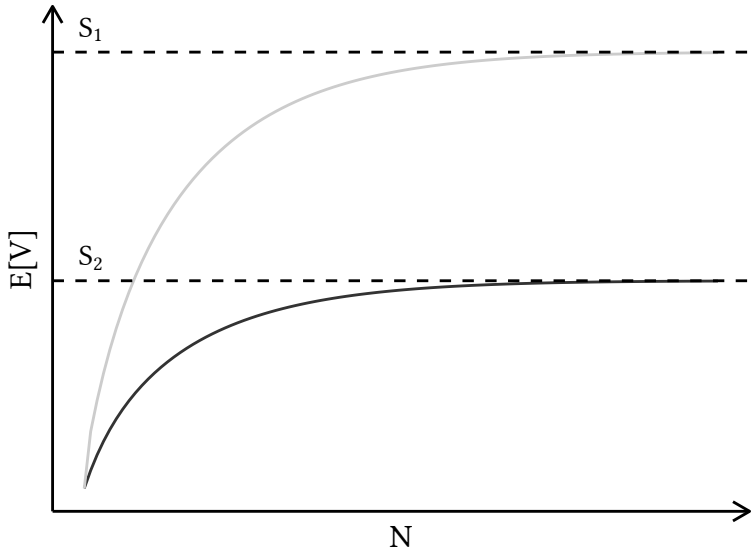

Figure 14: A schematic illustration of S for two different vocabulary growth curves. The curve with the greater S is considered more productive than the curve with the smaller S.

The idea of a maximum vocabulary size for all derivational processes, whether productive or unproductive, is also supported mathematically by the analyses that both Hartmann (2016) and I have run of the dependency of $\mathscr{P}$ on the sample size N. Recall that $\mathscr{P}$ is the slope of an interpolated VGC, and that as N gets arbitrarily large, $\mathscr{P}$ approaches zero. A slope of zero means that the curve in question is perfectly horizontal at that point, so the trend of $\mathscr{P}$ toward zero in all of our analyses suggests that, as N increases and the slope approaches zero, the VGC would approach being horizontal.

The horizontal asymptote that the VGC approaches is used by certain models of vocabulary growth such as the finite Zipf-Mandelbrot model (to be discussed in more detail below), and it is referred to as S. S is the y value that the vocabulary approaches as the sample size approaches infinity, i.e. $S = \lim_{N \to \infty} V$ (Zeldes 2012: 79, Baayen & Lieber 1991: 817). In linguistic terms, the value given by S is the maximum expected number of types that a morphological process would produce, given an arbitrarily large sample of tokens N (Baayen 2001, Zeldes 2012, Evert 2004).

The interpretation of S is simple: More types potentially created means more productivity. This is visualised in Figure 14, where the curve that extends farther up on the y axis represents the more productive process, since it would produce more types than the one extending less far.

Since we do not know how long it might take for any curve to start to flatten out, we have to generate a model of a VGC that can be extended to any arbitrary sample size. Models that do this are the LNRE (Large Number of Rare Events) models described

by Baayen (2001). LNRE models are said to overcome the dependency on sample or corpus size that has been a problem for the previously discussed measures, since smaller samples can just be extrapolated to be the same size as the largest one, or beyond (cf. Baroni & Evert 2014, Lüdeling 2009). As I will illustrate below, however, LNRE models also show a dependency on the original sample size.

The model that has been most widely used for quantifying morphological productivity is the finite Zipf-Mandelbrot (fZM) model by Evert (2004). In what follows, I will briefly explore how this model works, so that we can understand how the sample size dependency might come into play.

The basis of the model is the Zipf-Mandelbrot Law of word frequency distributions. This law describes the frequencies with which words appear in a sample. It is given in Equation 5, where z is the so-called Zipf rank of a word (the most frequent word has rank $z = 1$, the second-most frequent $z = 2$, and so on), f_z is the frequency of the word with rank z, C is a constant that is roughly equivalent to the frequency of the word with rank 1, and a and b are positive constants that can be used to tweak the shape of the curve at the upper and lower extremes.

$$f_z = \frac{C}{(z+b)^a} \tag{5}$$

A few words of background: the original law of word frequency distributions was Zipf's Law, which is simply $f_z = \frac{C}{z}$ (Zipf 1949). George Kingsley Zipf – "undoubtedly the father of lexical statistics" (Baroni 2009: 803) – was the first modern researcher to recognise that there is a systematic relation between a type's frequency and its rank. With increasing rank, the type's frequency decreases, very steeply at first and then increasingly shallowly. In other words, the frequency "decreases in a harmonic series: the second item is half as frequent as the first, the third is a third as frequent as the first, and so on" (Zeldes 2012: 77–78).

Zipf's Law is a general fact about language, and it holds independent of corpus, specific language, register, tokenization and type-mapping methods, text genre, etc. (in fact, it also shows up in all sorts of areas outside of language as well; Baroni 2009: 810). And as I will illustrate, Zipf's Law also describes the type frequency distribution of words derived with a morphological process (cf. Zeldes 2012: 78).

Mandelbrot's (1953) adaptation to Zipf's Law was the addition of the two parameters a and b. These parameters are there to adjust the upper and lower ends of the distribution, because Zipf's Law tends to overestimate the frequency of very frequent and very infrequent types (Mandelbrot 1953: 492). Adding the constant b to z in the denominator significantly increases the denominator's value when z is small (i.e. for the high-frequency types), which decreases the value of the overall expression compared to

Zipf's original model for the same z. The role of b becomes negligible as z gets larger (i.e. as we move toward the low-frequency types). But as z increases, the other new parameter a in the exponent of the denominator causes the denominator to again grow larger than in Zipf's original model, decreasing the overall value of the expression at the other end as well. In this way, these two parameters tidy up the extremes of the type frequency distribution (Mandelbrot 1953: 492, Zeldes 2012: 80, Baayen 2001: 101–102). Zipf's Law can be understood as a special case of the Zipf-Mandelbrot Law, where $a = 1$ and $b = 0$ (Zeldes 2012: 80).[5]

Rather than estimating the actual observed frequencies of words in the sample, which will be "subject to sampling error, and will therefore diverge slightly" from the values predicted by the Zipf-Mandelbrot Law in Equation 5 (Baayen 2001: 15), we can rephrase this law in terms of the probability of types, assuming as Baayen does that "[u]nderlying the observed frequencies, there is a distribution of probabilities for which Zipf's law should also be valid." After all, we are interested in how probable it is that we will encounter new types at any given size of N, not in how many of these types there might be. The Zipf-Mandelbrot Law is restated as its corresponding probability distribution in Equation 6, where π_z is the probability of the word with rank z (Baayen 2001: 15).

$$\pi_z = \frac{C}{(z+b)^a} \tag{6}$$

In the frequency-based expression in Equation 5, the role of C was to ensure that all frequencies summed up to the total number of tokens in the sample. In Equation 6, C's role is now as a normalising constant to ensure that all the probabilities sum up to one (Baayen 2001: 15, Zeldes 2012: 78).

At this point, let us pause and clarify why this is important to know. The model can best extrapolate VGCs when the observed frequency distribution in the sample is close to the probability distribution described by Equation 6. However, the smaller the samples are, the worse they will correspond to this frequency distribution, which is characterised by a few very frequent types and very many rarer types (which are the "large number of rare events" referenced by the model name). The larger the sample, the steeper the drop-off between the frequency of the top few types and the longer the tail (Zeldes 2012: 81). This is certainly apparent in my data from RIDGES; see Figure 15 for the frequency distributions of the top twelve most frequent types for each suffix, and notice that the

[5]In Baayen (2001: 14) and Baroni (2009: 813), a slightly different version of Zipf's Law is given, in which the denominator is already raised to the power of a. In Zipf's (1949) original equation, though, there was no exponent, and in Mandelbrot's adaption of Zipf's Law, he explicitly states that he introduces the exponent in the denominator, implying that it was not there initially (Mandelbrot 1953: 491–492). Baayen (2001: 16) mentions that Zipf often took a to equal unity, i.e. one, in which case its presence or absence in Zipf's Law doesn't affect the resulting value either way, but I have opted to present the equation for Zipf's Law following the originals in Zipf (1949) and Mandelbrot (1953), without the parameter a.

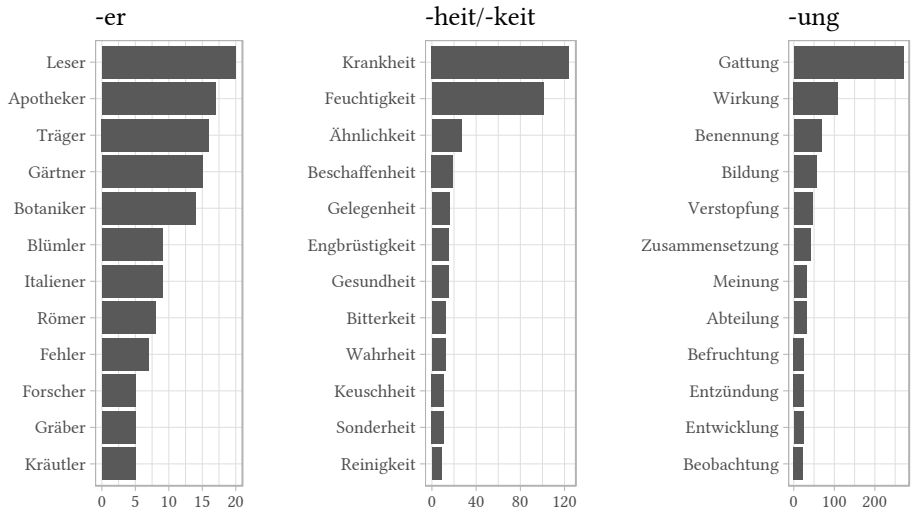

Figure 15: The twelve most frequent types for each suffix, illustrating that the steep decrease and long tail of the expected Zipfian distribution only starts to emerge as sample sizes get larger (*-ung*: $N = 277$; *-heit/-keit*: $N = 652$; *-ung*: $N = 1992$).

expected steep drop-off and long tail show up best in the large sample for *-ung* and worst in the small sample of *-er*.

These two facts together – deviation from the probability distribution used for the model leading to worse estimations, and smaller samples being more likely to diverge from this distribution due to sampling effects – suggest that the model estimation will generally be less successful for smaller samples than for larger ones. This could cause difficulties for the variably-sized and overall fairly small samples from historical corpora. Let us bear this in mind as we continue on.

To create an LNRE model that approximates a VGC from a probability distribution, one takes the first derivative of this distribution function to get the corresponding density function (cf. Evert 2004: Eqs. 15, 16). This density function resembles a VGC and can be tweaked to fit the observed data as closely as possible, using parameters available in the expression, such as the constants a and b. The result of applying this procedure to the probability-based expression in Equation 6 is the Zipf-Mandelbrot LNRE model (the ambitious reader is referred to Evert 2004, Baayen 2001 for more detail).

We are not quite at the final form of this model used for measuring productivity, though, since the simple Zipf-Mandelbrot model only describes an infinite vocabulary, e.g. randomly-generated character strings. A consequence of this is that the resulting VGC will always continue to grow, approaching infinity (Evert 2004: 6).

However, natural language (considered in finite chunks as discussed above) does not have an infinite vocabulary. Applying the Zipf-Mandelbrot model to natural language data results in the model greatly overestimating the observed values at larger sample

sizes, since the curve continues ever upward after the vocabulary of natural language has eventually been exhausted (Evert 2004). So, an adaptation to this model is necessary.

To create the *finite* Zipf-Mandelbrot model, Evert (2004: 7) modifies the original Zipf-Mandelbrot model by adding a "lower cut-off point $A > 0$ for the type density" that the probability π of any given item may not subceed. In plainer terms, this restriction means that there may be no types with probability $\pi < A$ in the vocabulary. Placing a lower bound on the probability in this way means that the model no longer predicts infinitely many low-probability – i.e. rare – types, but only a finite number of them. Consequently, at some point, the modelled density curve will reach a horizontal asymptotic limit, since eventually, the threshold of the lowest-probability, rarest type that the model predicts will be reached. This horizontal asymptote that the VGC approaches is our familiar S. Evert (2004: 10) validates his fZM model, showing that it convincingly approximates unseen data and fits natural language data better than the non-finite Zipf-Mandelbrot model can.

This does seem hopeful at first blush. Based only on a frequency spectrum from the observed data, the fZM model can extrapolate a VGC to find S, the maximum number of types we would encounter if the number of tokens in our sample could be arbitrarily large. Because the principle of the model is to extrapolate to arbitrary sample sizes, rather than focus only on observed data, the value of S should in theory be directly comparable, no matter the original corpus size (cf. Baayen 2001, Tweedie & Baayen 1998). As I suggested above, though, it turns out that this is not quite so.

Previous work has shown that most so-called constants in lexical statistics aren't really constant, but show some systematic dependency on sample size N. Baayen & Tweedie (1998) observed already two decades ago that "[t]he parameters of LNRE models are in theory invariant with respect to the sample size" but that "in practice the parameters of LNRE models may nevertheless reveal substantial dependence on N" (Baayen & Tweedie 1998: 145). Tweedie & Baayen (1998) verified this for several so-called constants from older LNRE models, but to my knowledge, S has not yet been inspected for this property. Showing that S is indeed also dependent on sample size is my goal for the following section.

4.4.1 The constancy of S

There is already reason to believe that S may also show a systematic dependency on the number of tokens in the sample. Baroni & Evert (2014: 13) mention an "undesirable dependency of LNRE model estimation on sample size" in their illustration of an fZM model for the Italian prefix *ri-*. They show that an fZM model with $N = 1{,}399{,}898$ resulted in $S = 78{,}194{,}057$, but when they calculated the same model on only a subset of their data

with $N = 700{,}000$, the resulting S was much lower at $S = 32{,}353{,}911$ (Baroni & Evert 2014: 10, 13).[6]

To try to detect such a dependency from the RIDGES data, I followed the same procedure as outlined in Section 4.2.1 above to calculate Monte Carlo means and confidence intervals for S at twenty equally-spaced measurement points with progressively increasing sample sizes for each suffix *-er*, *-heit*/*-keit*, and *-ung*.[7]

I computed 2,000 iterations here as well, but for two reasons I did not have 2,000 data points available for analysis at each measurement point. On the one hand, sometimes the fZM model's parameter estimation failed, giving no result for S at that point (cf. Baroni & Evert 2014: 12). On the other, sometimes the model predicted inconceivably large values for S (occasionally on the order of 10^{55}; cf. Schneider-Wiejowski 2011: 188, who found the same thing), so I also removed all outliers from the distribution at each measurement point. The number of model estimation failures and outliers for each suffix at each measurement point are provided in Appendix B.

After calculating the Monte Carlo distributions at each measurement point, I cut down all data sets for each suffix to the maximum common size, randomly discarding the excess data. This meant that the Monte Carlo means and 95% confidence intervals were always computed over an equal number of data points.

The resulting plots are presented in Figure 16. The numbers of iterations used to compute the data at each measurement point are given in the y axis label. Here, as in Figure 8 (the parallel figure for $\mathscr{P}$), we see the confidence intervals getting smaller as the samples' homogeneity increases, i.e. as the random selections out of the maximum N tokens resemble one another more and more.

Again, what we are interested in is whether a horizontal line would be compatible with this confidence range, since we would expect a horizontal line if S were constant over all N. None of these plots allow us to definitively discount this, because the sample sizes are so small. The first two plots for *-er* and *-ung* definitely do not disprove it, but the confidence intervals for *-ung* (the largest sample) are starting to suggest an upward trend.

Because the exploration based on the RIDGES data was inconclusive, I also carried out the same procedure on a much larger dataset of 10,000 tokens of Modern German *-heit*/*-keit*, gathered from the DECOW16B web corpus (Schäfer 2015, Schäfer & Bildhauer 2012). I chose *-heit*/*-keit* because it is the most formally distinctive of the three

[6] Besides the dependency on N, it is unclear how to interpret these S values linguistically. It seems implausible that there might be tens of millions of types derived with this morpheme when the number of total types in a language is generally understood as several degrees of magnitude smaller. Even English, with its very large vocabulary, is estimated only to have around 750,000 types (Lexico.com n.d.).

[7] The options I used for the fZM model were the defaults, with approximations allowed (see Evert & Baroni 2017 for more details about customising the model).

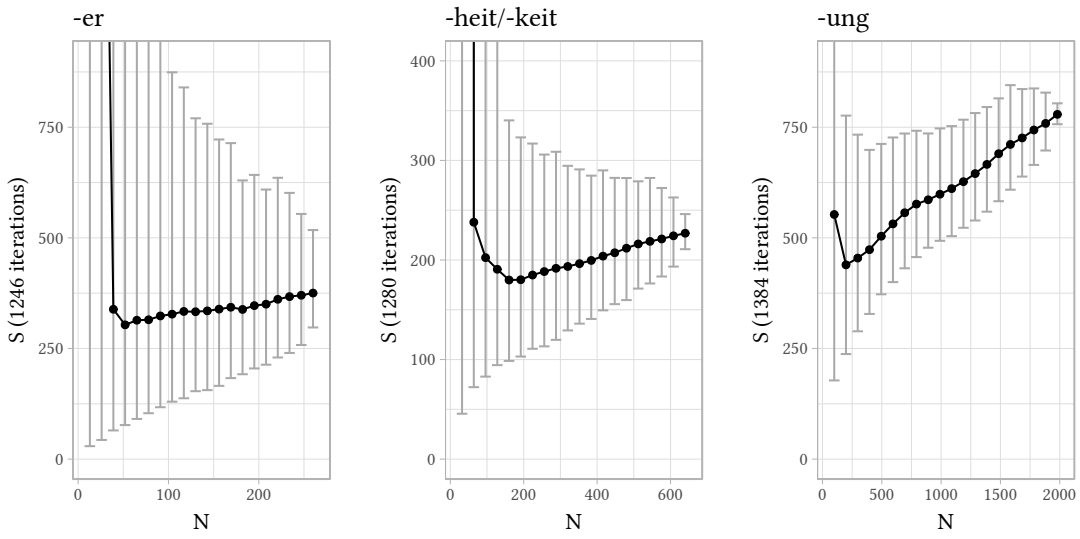

Figure 16: Monte Carlo means and 95% confidence intervals for S for *-er*, *-heit*/*-keit*, and *-ung* in RIDGES.

suffixes, so the sample will contain the fewest false matches compared to *-ung* and *-er*. I lemmatised the sample of 10,000 tokens following the same guidelines that I used for the RIDGES data (see Section 3.3) and conducted the Monte Carlo analysis in the same way as above, except that I used 100 measurement points instead of twenty. This was done to keep the step size of 100 tokens between measurement points more comparable to the step sizes in the RIDGES samples (11 for *-er*, 32 for *-heit*/*-keit*, and 99 for *-ung*). The resulting plot of the Monte Carlo means and 95% confidence intervals for *-heit*/*-keit* from DECOW16B is given in Figure 17.[8]

From this plot, we see that no horizontal line is compatible with the confidence range. This shows clearly that S is not constant over all sample sizes, but increases as the sample size N increases. We can therefore conclude that S shows a systematic dependency on the size of the sample N used to compute the fZM model, in line with the findings from Tweedie & Baayen (1998) about other putative constants from different LNRE models.

One source of this dependency might be the above-mentioned fact that LNRE models work less well when the frequency distribution in the data diverges greatly from the expected probability distribution that underlies the model. As we saw, this is more likely to happen in smaller samples. So, the type frequency distributions becoming more and more typically Zipfian as the sample gets larger – with a steeper drop-off from the highest-frequency types and a longer tail among the low-frequency ones (Zeldes 2012:

[8] At the final measurement point, the sample of 10,000 was completely exhausted, so the data used each time was identical. This is why both the upper and lower bounds of the confidence interval are the same as the mean. In contrast, the RIDGES samples are not evenly divisible by twenty, which is why they still show a little variability at the final measurement point: the samples were never fully used up like this one was.

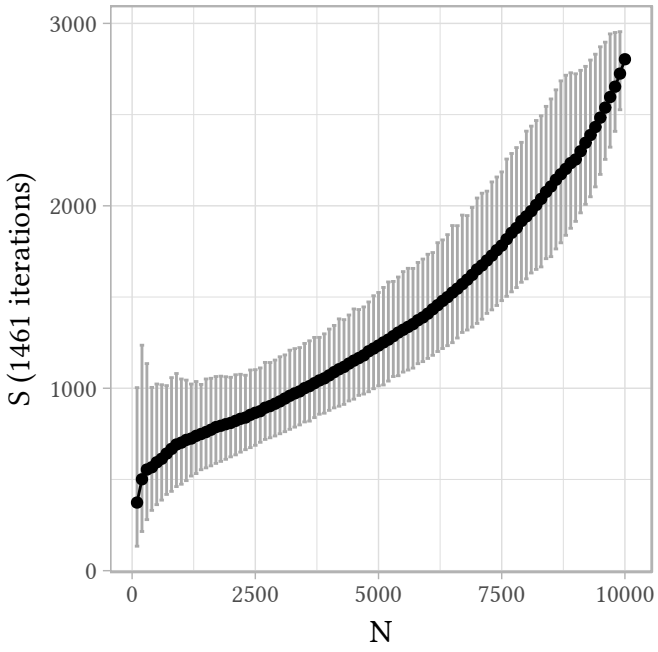

Figure 17: Monte Carlo means and 95% confidence intervals for S for a sample of 10,000 *-heit*/*-keit* tokens from DECOW16B, measured at 100 points, each 100 tokens apart.

81) – could affect the way that the model estimates S. That said, a complete understanding of the mathematics involved in the model estimation would be required to determine the actual source of the dependency.

4.4.2 The randomness assumption and natural language data

Before moving on to the summary and discussion, I want to focus on another problematic aspect of LNRE models, namely their assumption that the distribution of words in texts is random. This is also referred to as the urn model, or in more current parlance, the bag of words model (Johnson & Kotz 1977, Tweedie & Baayen 1998). Using a word is modelled the same way as taking a marble of a particular colour out of an urn that contains many differently-coloured marbles, where the colour of the marble just withdrawn has no effect or influence on the colour of the marble that follows next.

This model is problematic because it does not reflect actual language use. Words are actually highly interconnected and interdependent, but this assumption is used in LNRE models to simplify the mathematics (Tweedie & Baayen 1998: 349).

Predictably, this means that the model's behaviour on randomised data (the data it is designed to deal with) diverges from its behaviour for natural language data, which behaves more variably. For example, in natural language we might see clustering of rare, productively formed words, a phenomenon called "underdispersion" (cf. Zeldes 2012: 84). This happens because a term that has been spontaneously coined for the speaker or

writer's current communicative needs will often be repeated a few times within a short amount of text (and then possibly never again). Consider the following dialogue from Ward (2018):

A: "Back in those days it was more like two million dogs a year were euthanised because they couldn't find homes in the shelter systems. [...] It's debatable too because shelters are not required to keep numbers on their euthanisations, so it's all an average."
B: "Oh, I thought that they would have to have spreadsheets!"
A: "No, not the euthanisations, because some shelters euthanise so fast."

In Speaker A's first utterance, the noun *euthanisation* was derived to talk about the act of euthanising, and then reused in their second utterance to refer to the same, still-relevant concept. The recurrence of rare words in such a short span of text is not accounted for by a model that assumes that words are randomly distributed in a text. Because these words occur so closely together, the model understands them to be highly frequent, leading it to overestimate type counts for rare words (cf. Zeldes 2012: 84, Baayen 2001, Evert 2004: 10–11, Mandelbrot 1961: 218).

We can tell that the model behaves differently on randomised data than it does on natural language data if we overlay the empirical values for $\mathscr{P}$ and S from the RIDGES samples (i.e. from natural language data) onto the plots with the Monte Carlo means and confidence intervals (i.e. from randomised data). As we see in Figure 18, most of the empirical natural language values lie outside of the Monte Carlo confidence ranges. This result harmonises with Tweedie & Baayen (1998: 349), who also find that, "[w]hen the empirical values of the text constants are compared with the theoretical values, they frequently fall outside the 95% MC [Monte Carlo -EP] confidence limits established" (Tweedie & Baayen 1998: 349)

We also see that all but one of the empirical values (S for *-ung* in the subperiod 1482–1549) are below the means of the Monte Carlo distributions. This probably reflects the way that these models of vocabulary growth, both interpolated (in the case of $\mathscr{P}$) and extrapolated (in the case of S), overestimate natural language values.

In sum, what these plots clearly show is that the measures $\mathscr{P}$ and S behave differently for the natural language data that they are *intended* to model than for the randomised data that they are *designed* to model. By using these models, we are implicitly asked to accept the assumption that language is random, but given the striking divergence of the results for empirical and randomised data, I would argue that we should not do so.

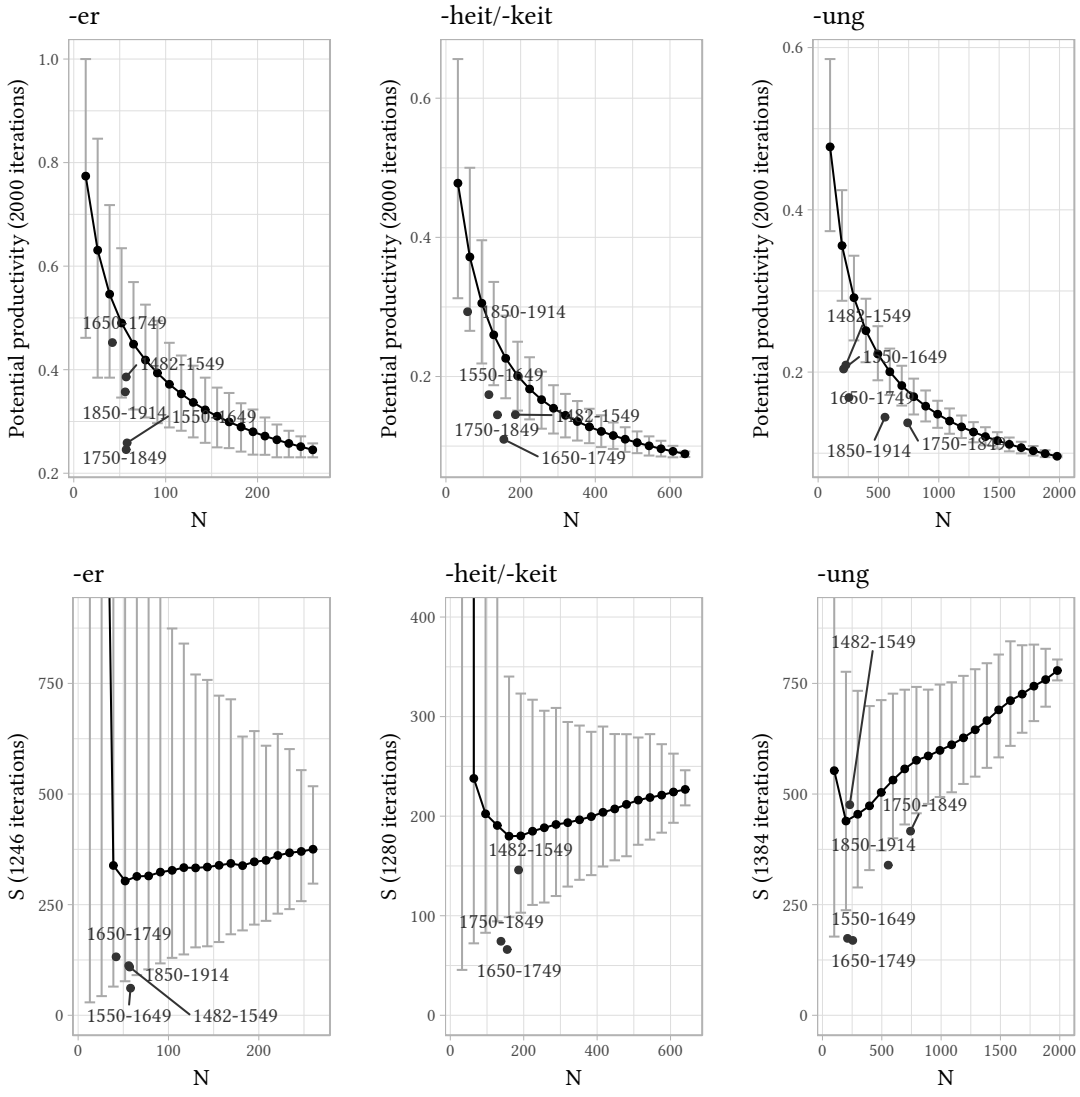

Figure 18: Monte Carlo means and 95% confidence intervals for $\mathscr{P}$ (top row) and S (bottom row) for each suffix, including the empirical data points for each suffix's sample from the indicated subperiod. (Missing data points for S are instances of model estimation failure.)

4.5 Summary and outlook for existing measures

This section has explored the primary measures that have been used in the literature to quantify different facets of morphological productivity and compare how it may change over time. I will briefly summarise these findings here and ultimately argue that, even though we can use certain sampling techniques to overcome the sample size dependency of measures like $\mathscr{P}$ and S, their behaviour with small samples still renders them of little use for our purposes.

Type counts, new type counts, and normalised type counts have been used as a reflection of past productivity of a process, since the number of different words that a derivational process has created reflects how widely the process has been applied in the past. However, type counts cannot be blindly compared without considering the number of tokens that contain them, and they should also always be understood in relation to the size of the text or subcorpus, since a larger subcorpus will contain more tokens and thus potentially more types.

In an attempt to overcome problems with differently-sized subcorpora, normalisation is sometimes applied to type counts so that they are understood as n types out of, say, 10,000 words in the corpus. However, normalising these counts introduces a decreasing dependency of type count on original corpus size, which means that normalised values – while apparently scaled the same way – do not seem to be sensibly comparable either. That said, this dependency might actually effectively mask the effects of the larger texts containing more types by making the normalised type count relatively smaller. Future work might look at developing a new normalisation procedure that explicitly takes all of this into account.

Hapax legomena are often used to approximate neologisms, the idea being that words that have just been newly formed will not have been used before in the sample and can be detected by sampling words that only appear once. However, the connection between neologisms and hapax legomena is tenuous in the small samples that are part and parcel of historical corpus work. Not all neologisms that really existed will be captured by the hapax legomena in the sample, nor will all the hapax legomena in the sample be neologisms, since the smaller the sample, the greater the chance is that non-neologisms will be sampled only once.

Vocabulary growth curves (VGCs) plot the way the type-token ratio changes as we proceed token by token through a sample. They are a useful tool because the VGCs of differently productive processes will look different: the more productive a process is, the more types it is likely to create, so the higher it will grow before eventually flattening out. However, when it comes to small or variously sized samples, VGCs are difficult to compare, because the part of the curve that we see is only the beginning, and we cannot

discern the shape of the ultimate curve from the small amount of data that we have.

Models exist to smooth out the wobbly empirical VGCs up to the observed sample size (interpolated VGCs) or to extend the existing data to arbitrarily large sample sizes, to see how the VGC would likely continue (extrapolated VGCs). Potential productivity $\mathscr{P}$ is the slope at the endpoint of an interpolated VGC, indicating how steeply the curve is growing at that point, i.e. the rate at which new types are being encountered. $\mathscr{P}$ can also be understood as a probability, based on the number of hapaxes out of the total number of tokens seen in the sample: it tells us how likely it is that the next token to be encountered is a hapax legomenon (assumed to be a good stand-in for a newly-formed word). However, $\mathscr{P}$ shows a problematic dependency on the number of tokens in the sample, such that the values for $\mathscr{P}$ from two samples with different numbers of tokens cannot sensibly be compared. After all, these reflect VGCs that end at different points, and we do not know how the curves will continue. Thus $\mathscr{P}$ does not tell us very much about the sample's ultimate VGC, so it is difficult to actually conclude anything about a morphological process' potential using this measure.

To overcome the problem of not knowing how VGCs continue beyond observed sample sizes, LNRE models are used: they model the observed data and extrapolate this model farther than the available number of tokens. The most widely used LNRE model in productivity studies is the finite Zipf-Mandelbrot (fZM) model. This is probably because it is easy to apply, thanks to the `zipfR` package in R, and also easy to interpret by comparing the resulting parameter S, which is a so-called constant that represents the maximum number of types a morphological process would create given an arbitrarily large sample of tokens. I have shown that S is not actually constant, but gets larger as the size of the sample over which the model is calculated increases. Also, LNRE models are built around an assumption that language is random, which simplifies the mathematics considerably, but falls short of reflecting the way that language is actually used. One must be willing to accept this assumption in order to use these models, and given the strong divergence of natural language data from randomised data, perhaps this assumption should not be made.

If, for whatever reason, one still wants to use $\mathscr{P}$ and S, the number of tokens in each sample has to be the same in order to sensibly compare the resulting values. There are several ways to achieve this. I do not recommend this, but I will mention these techniques here for completeness' sake.

The first approach involves taking subperiods of a set number of years and whittling each of the corresponding subcorpora down so that they all contain the same number of tokens, throwing out data until the largest common sample size is reached (cf. Zeldes 2012: 65, 70–71). As Baayen (2001: 8) points out, though, there is no principled way to

decide which data to leave out, and this approach also overvalues regular subperiods, which are just an idealisation that might even obscure the way the data really changes over time (cf. Claridge 2008). The second approach sacrifices strict temporal divisions in order to minimise how much data is discarded. This method involves dividing the corpus so that each subcorpus contains the same number of tokens containing the morpheme of interest, regardless of time period (Dal & Namer 2016: 75, Gaeta & Ricca 2003). Some of these subcorpora will span longer subperiods and some shorter, but the token count will always be the same, so that the measures can be compared across all subcorpora. This approach maximises the amount of data that can be evaluated, though it still falls prey to the idealisation of chunking time up into mutually exclusive subperiods.

Another method of maintaining the same token counts in each sample and at the same time overcoming the problems of subperiodisation would be to employ sliding token-based windows. We would arrange the sample in chronological order, set a window length of, say, 200 tokens, and calculate the measures over all tokens in that first window. Then this window could be shifted over by, say, 25 tokens, so that the first 25 tokens in the sample are removed from the window and the next 25 unseen tokens are added, which means that 150 of the tokens stay the same between the two windows each time. This is a more accurate reflection of time as a continuous variable.

While sampling in this way would meet the requirement of equal sample sizes, we have to confront the fact that $\mathscr{P}$ and S are not particularly useful for small samples. Especially for historical samples, which are small to begin with, in order to track any kind of diachronic development the windows need to be smaller still. This creates problems both for the interpretation of $\mathscr{P}$ and for the estimation of S.

For a small sample, a VGC will end quite close to where it begins, and near the beginning, every curve is growing steeply. We are thus more likely to get higher values of the slope $\mathscr{P}$, the smaller the sample, which would suggest an unrealistically high degree of productivity. And we encounter again the problem that in smaller corpora, hapax legomena – central to the interpretation of $\mathscr{P}$ as a probability telling us about how likely we are to encounter neologisms – approximate neologisms less well (cf. Cowie 1999, Cowie & Dalton-Puffer 2002).

Turning now to S: The fZM model's parameter estimation fails more frequently and produces more outliers for smaller samples than for larger ones (see the numbers for the Monte Carlo analysis of S in Appendix B). The sample size that is needed for the Poisson random sampling model used for all LNRE models in `zipfR` to be reliable is minimum 10,000, which is far outside the realm of feasibility for sliding windows over historical corpus samples. An option for multinomial sampling, which should be more accurate for smaller samples, is apparently under development but has not yet been

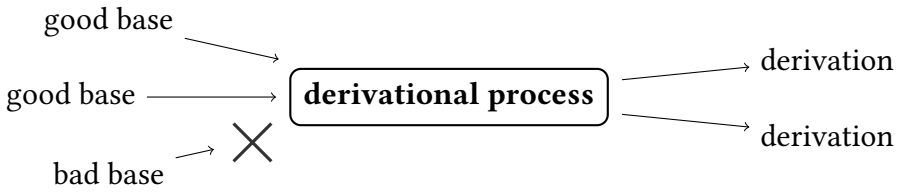

Figure 19: An abstract schema of a derivational process that accepts some (but not all) bases as input and produces derived words as output (adapted from Figure 1).

implemented (Evert & Baroni 2017: 20). And in any case, these models have been shown to overestimate expected values compared to the actual data, so they do not offer a very accurate reflection of natural language use.

In sum, the main take-away of this section is rather negative. The measures most widely used to try to quantify morphological productivity and compare changes in it over time are not truly comparable between differently-sized subcorpora, and while sampling techniques exist to create equally-sized samples, these would still be too small for the measures to be sensibly interpreted. This means that the current method of quantifying diachronic productivity – computing some measures of productivity over corpus data across several subperiods and comparing the results – is fundamentally flawed. In the following section, I will suggest an alternate approach to studying changes in productivity over time that might be a good next step for future research in this area.

5 Another way forward

My exploration of the behaviour of productivity measures on small, diachronic samples has revealed the unfeasibility of the usual methodology for quantifying changes in morphological productivity over time. We can put forward some desiderata for what the ideal measure of diachronic productivity should do – take into account original corpus size, be comparable across different sample sizes, reflect actual natural language use rather than the marbles-from-an-urn assumption, etc. – but perhaps the real question is not what this sort of measure should do, but whether we should even be trying to measure productivity in this way.

Early on, I introduced a schema for how we can conceptualise derivational processes, reproduced here in more abstract terms in Figure 19. I discussed how the current quantificational approach is firmly settled on the output side of the schema. With this approach, we analyse a sample of all words containing the morpheme of interest, without considering any characteristics of the process itself. However, if we come back to what we know about productivity, we see that there are fundamental conceptual problems with looking only at the process' output and not considering its input restrictions.

Crucially, our intuition about the productivity of a process requires an understanding of the process' use *within that process' set of possible inputs*. To illustrate: a process can have very restrictive input constraints, leading to a very small set of inputs, and yet it can still be used very freely within that input set. For example, in Modern German, the prefix *ein-* can be applied to any verb that means 'to sleep' to create a verb meaning 'to fall asleep': *schlafen* → *einschlafen*, *schlummern* → *einschlummern*, *pennen* → *einpennen*, and so on. Also, the French suffix *-u* has been said to readily create adjectives of inalienable possession from any noun denoting a body part, e.g. *barbe* 'beard' → *barbu* 'bearded' (Dal & Namer 2016: 73).

Now, if we were to create VGCs for *ein-* and *-u* (in so doing, ignoring the restrictions on the processes' input sets), we would find only a small number of different types for each because the input sets are so small. This would lead us to conclude that each process is very unproductive, even though it is actually very productive within that extremely small set of words.

These examples illustrate the weakness of purely output-oriented measures of productivity like the ones widely used in the diachronic productivity literature. Quantitative methods should somehow reflect our intuitions or cognitive reality, so that we can learn something from them about how we actually use language (cf. Milin et al. 2016, Baayen et al. 2011). This is something that the measures discussed above do not do. We must return to what we know about how morphological productivity works in order to find a suitable method for measuring it.

One thing that we know, as mentioned at the beginning, is that speakers have strong intuitions about which affixes are possible and which are not. For example, Bauer (2001: 67) discusses a conversation about deriving a noun meaning 'the quality of being a chair', where *chairness*, *chairity*, and *chairosity* were rejected before the group finally settled on *chairhood*. What we learn from examples like this is that, when a base word needs to be derived, each derivational morpheme has a particular likelihood of being applied, given certain linguistic and extralinguistic factors (cf. Cowie & Dalton-Puffer 2002: 412).

To determine the probability that each morphological process has of applying to the given base word, we must consider the ways in which these linguistic and extralinguistic factors shape the potential input to a process. This can be done with statistical methods like regression, which can show which factors play a role in predicting derivation of an input word with a particular suffix. These probabilities reflect our natural intuitions as language users and our reasoning processes when trying to productively form a word based on the characteristics of the base.

Using regression in historical corpus linguistics is not unheard of. For example, Wolk et al. (2013) use logistic regression to model various predictors affecting dative and gen-

itive alternations in Late Modern English. They examined how the roles of the relevant factors changed over time by including "real time" as a predictor and examining its interactions with the other predictors (cf. Wolk et al. 2013: 403). This "real time" predictor is a good reflection of time as a continuous variable, because it does not involve any subperiodisation. Instead, it measures the chronology of each text in centuries since 1800, where 1800 is approximately the midpoint of their sample, "so that a text from 1651 would count as (1651 − 1800) / 100 = −1.49 and a text from 1931 as (1931 − 1800) / 100 = 1.31" (Wolk et al. 2013: 394).

Applying this or similar regression methods to the question of how a derivational process' input restrictions change over time seems like a more promising way forward than continuing to pursue these flawed productivity measures.

6 Conclusion

In this thesis, I have explored the primary way in which historical linguists try to quantify the change in morphological productivity over time. This method uses type counts to assess the past extent of use of a morphological process until the time of sampling, as well as measures like $\mathscr{P}$ and S to estimate the potential of a process to create new forms. These measures are calculated over several subperiods and then compared to draw conclusions about how productivity changes over time.

I have shown that not only can these measures not be used to accurately compare differently-sized samples, they also do not reflect language users' intuitions about how morphological productivity really works. Crucially, the measures do not consider important facets of productivity like the size of a morphological process' set of possible inputs. Analyses of a process' productivity must always involve an understanding of its input restrictions, because these are an important part of speakers' linguistic competence when they use derivational morphology. Statistical methods like regression may be better suited to studying changes in morphological productivity than the present measures available. In any case, a different way forward, one that reflects and tells us something about how we as language users understand morphological productivity, is necessary.

References

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: MIT Press.

Baayen, R. Harald. 2001. *Word frequency distributions*. Dordrecht/Boston/London: Kluwer Academic Publishers.

Baayen, R. Harald. 2003. Probabilistic approaches to morphology. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 229–287. Cambridge: MIT Press.

Baayen, R. Harald. 2009. Corpus linguistics in morphology: morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 900–919. Berlin: De Gruyter.

Baayen, R. Harald & Rochelle Lieber. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29. 801–844.

Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438–481.

Baayen, R. Harald & Fiona J. Tweedie. 1998. Sample-size invariance of LNRE model parameters: Problems and opportunities. *Journal of Quantitative Linguistics* 5(3). 145–154.

Baroni, Marco. 2009. Distributions in text. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 803–822. Berlin: De Gruyter.

Baroni, Marco & Stefan Evert. 2014. *The zipfR package for lexical statistics: A tutorial introduction*. Available from http://zipfr.r-forge.r-project.org/.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge/New York/Melbourne: Cambridge University Press.

Belz, Malte, Carolin Odebrecht, Laura Perlitz, Gohar Schnelle & Vivian Voigt. 2018. *Dokumentation und Annotationsrichtlinien für das Korpus RIDGES Herbology Version 8.0*. Humboldt-Universität zu Berlin.

Booij, Geert. 2012. *The grammar of words: an introduction to linguistic morphology*. 3rd edn. Oxford: Oxford University Press.

Brendel, Bettina, Regina Frisch, Stephan Moser & Norbert Richard Wolf. 1997. *Wort- und Begriffsbildung in frühneuhochdeutscher Wissensliteratur: Substantivische Affixbildung* (Wissensliteratur im Mittelalter 26). Wiesbaden: Reichert.

Claridge, Claudia. 2008. Historical corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, Berlin. 242–259: Mouton De Gruyter.

Cowie, Claire. 1999. *Diachronic word-formation: A corpus-based study of derived nominalizations in the history of English*. University of Cambridge dissertation.

Cowie, Claire & Christiane Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In Javier E. Díaz Vera (ed.), *A changing world of words: Studies in English historical lexicography, lexicology and semantics*, 410–437. Amsterdam: Rodopi.

Dal, Georgette & Fiammetta Namer. 2016. Productivity. In Andrew Hippisley & Gregory T. Stump (eds.), *The Cambridge Handbook of Morphology* (Cambridge Handbooks in Language and Linguistics), 70–89. Cambridge: Cambridge University Press.

Demske, Ulrike. 2000. Zur Geschichte der ung-Nominalisierung im Deutschen: Ein Wandel morphologischer Produktivität. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)* 122(3). 365–411.

Doerfert, Regina. 1994. *Die Substantivableitung mit -heit, -keit, -ida, -î im Frühneuhochdeutschen.* Berlin: De Gruyter.

Evert, Stefan. 2004. A simple LNRE model for random character sequences. In *Proceedings of JADT 2004: 7es Journées internationales d'Analyse statistique des Données Textuelles.*

Evert, Stefan & Marco Baroni. 2007. zipfR: word frequency distributions in R. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, 29–32.

Evert, Stefan & Marco Baroni. 2017. *Package 'zipfR'*. R package documentation.

Evert, Stefan & Anke Lüdeling. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In *Proceedings of Corpus Linguistics 2001*, vol. 168.

Gaeta, Livio & Davide Ricca. 2003. Italian prefixes and productivity: A quantitative approach. *Acta Linguistica Hungarica.* 89–108.

Habermann, Mechthild. 2001. *Deutsche Fachtexte der frühen Neuzeit: Naturkundlich-medizinische Wissensvermittlung im Spannungsfeld von Latein und Volkssprache* (Studia Linguistica Germanica 61). Berlin / New York: De Gruyter.

Hartmann, Stefan. 2016. *Wortbildungswandel: Eine diachronie Studie zu deutschen Nominalisierungsmustern* (Studia Linguistica Germanica 125). Berlin / Boston: De Gruyter.

Hartweg, Frédéric & Klaus-Peter Wegera. 2005. *Frühneuhochdeutsch. Eine Einführung in die deutsche Sprache des Spätmittelalters und der frühen Neuzeit.* Tübingen: Max Niemeyer Verlag.

Johnson, Norman Lloyd & Samuel Kotz. 1977. *Urn models and their application; an approach to modern discrete probability theory.* New York: John Wiley & Sons.

Kempf, Luise. 2016. *Adjektivsuffixe in Konkurrenz: Wortbildungswandel vom Frühneuhochdeutschen zum Neuhochdeutschen* (Studia Linguistica Germanica 126). Berlin / Boston: De Gruyter.

Klein, Wolf Peter. 2010. Die deutsche Sprache in der Gelehrsamkeit der frühen Neuzeit. Von der lingua barbarica zur HaubtSprache. In Herbert Jaumann (ed.), *Diskurse der Gelehrtenkultur in der Frühen Neuzeit. Ein Handbuch*, 465–516. Berlin/New York: De Gruyter.

Krause, Thomas & Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities* 31(1). 118–139.

Lexico.com. N.d. *How many words are there in the English language?* https://www.lexico.com/en/explore/how-many-words-are-there-in-the-english-language. Accessed 04/07/2019.

Lüdeling, Anke. 2009. Carmen Scherer, Wortbildungswandel und Produktivität. *Beiträge zur Geschichte der deutschen Sprache und Literatur (PBB)* 131(2). 333–339.

Lüdeling, Anke, Stefan Evert & Ulrich Heid. 2000. On measuring morphological productivity. In *KONVENS 2000/Sprachkommunikation, Vorträge der gemeinsamen Veranstaltung 5. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS), 6. ITG-Fachtagung Sprachkommunikation*, 57–61. VDE-Verlag GmbH.

Lüdeling, Anke, Carolin Odebrecht, Laura Perlitz & Amir Zeldes. N.d. *RIDGES-Herbology (Version 8.0).* Humboldt-Universität zu Berlin. http://korpling.org/ridges/.http://hdl.handle.net/11022/0000-0007-C6A3-1.

Mandelbrot, Benoit. 1953. An informational theory of the statistical structure of language. *Communication theory* 84. 486–502.

Mandelbrot, Benoit. 1961. On the theory of word frequencies and on related Markovian models of discourse. In Roman Jakobson (ed.), *Structure of language and its mathematical aspects*, 190–219. Providence, Rhode Island: American Mathematical Society.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4). 507–526.

Müller, Peter O. 1993. *Substantiv-Derivation in den Schriften Albrecht Dürers: Ein Beitrag zur Methodik historisch-synchroner Wortbildungsanalysen*. Berlin / New York: De Gruyter.

Odebrecht, Carolin, Malte Belz, Amir Zeldes, Anke Lüdeling & Thomas Krause. 2017. RIDGES Herbology: designing a diachronic multi-layer corpus. *Language Resources and Evaluation* 51(3). 695–725.

Plag, Ingo. 1999. *Morphological productivity. Structural constraints in English derivation*. Berlin/New York: Mouton de Gruyter.

Plag, Ingo, Christiane Dalton-Puffer & Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Renze, John & Eric W. Weisstein. N.d. *Intermediate Value Theorem*. From MathWorld – A Wolfram Web Resource. http : / / mathworld . wolfram . com / IntermediateValueTheorem.html.

Rössing-Hager, Monika. 1990. Leitprinzipien für die Syntax deutscher Autoren um 1500: Verfahrensvorschläge zur Ermittlung zeitspezifischer Qualitätsvorstellungen, ihrer Herkunft und Verbreitung. In Anne Betten (ed.), *Neuere Methoden der historischen Syntaxforschung. Referate der Internationalen Fachkonferenz Eichstätt* (Reihe Germanistische Linguistik 103), 406–421. Tübingen: De Gruyter.

Ruh, Kurt. 1956. *Bonaventura deutsch. Ein Beitrag zur deutschen Franziskaner-Mystik und -Scholastik* (Bibliotheca Germanica 7). Bern: Narr.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*. Lancaster: IDS.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul, Turkey: European Language Resources Association (ELRA).

Scherer, Carmen. 2005. *Wortbildungswandel und Produktivität: Eine empirische Studie zur nominalen -er-Derivation im Deutschen*. Tübingen: Niemeyer.

Scherer, Carmen. 2007. The role of productivity in word-formation change. In Joseph C. Salmons & Shannon Dubenion-Smith (eds.), *Historical Linguistics 2005: Selected papers from the 17th International Conference on Historical Linguistics (Current issues in linguistic theory 284)*. Amsterdam/Philadelphia: John Benjamis Publishing Company.

Scherer, Wilhelm. 1878. *Zur Geschichte der deutschen Sprache*. 2nd edn. Berlin: Weidmann.
Schneider-Wiejowski, Karina. 2011. *Produktivität in der deutschen Derivationsmorphologie*. Bielefeld: Universität Bielefeld dissertation.
Spencer, Andrew. 2016. Two morphologies or one? Inflection versus word-formation. In Andrew Hippisley & Gregory Stump (eds.), *Cambridge Handbook of Morphology*, 27–49. Cambridge: Cambridge University Press.
Tweedie, Fiona J. & R. Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32(5). 323–352.
Ward, Alie. 2018. *Cynology: The study of dogs with Brandon McMillan*. Ologies (audio podcast). https://www.alieward.com/ologies/cynology.
Wegera, Klaus-Peter & Heinz-Peter Prell. 2000. Wortbildung des Frühneuhochdeutschen. In Werner Besch, Anne Betten, Oskar Reichmann & Stefan Sonderegger (eds.), *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung*, 2nd edn., vol. 2, 1594–1605. Berlin/New York: De Gruyter.
Wolk, Christoph, Joan Bresnan, Anette Rosenbach & Benedikt Szmrecsanyi. 2013. Dative and genitive variability in Late Modern English: exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.
Zeldes, Amir. 2012. *Productivity in argument selection. From morphology to syntax*. Berlin/Boston: De Gruyter Mouton.
Zipf, George K. 1949. *Human behaviour and the principle of the least effort. An introduction to human ecology*. New York: Hafner.

Appendices

A Corpus queries and export information

The following are the AQL queries used to gather the three samples for *-er*, *-heit*/*-keit*, and *-ung* from RIDGES. For *-heit*/*-keit* and *-ung*, an initial sample was conducted to detect what sort of graphematic variation appears on *dipl* and *clean*. The basis samples are the main samples used for analysis, and any graphematic variants or further forms that were not already captured in that sample were then added to it to create the full sample that was analysed (see Section 3.1 above for more detail).

-er	
Basis sample	`clean=/.*er/ _i_ pos=/[N].*/ _o_ norm`
Dative plural	`clean=/.*ern/ _o_ norm`

-heit*/*-keit	
Initial sample	`norm=/.*(k|h)eit(en)?/ _o_ clean _=_ dipl`
Basis sample	`norm=/.*(k|h)eit(en)?/ _o_ clean`
Graphematic variants	`clean=/.*heyt(en)?/ _i_ norm!=/.*(h|k)eit(en)?/`
	`clean=/.*keyt(en)?/ _i_ norm!=/.*(h|k)eit(en)?/`
	`clean=/.*hait(en)?/ _i_ norm!=/.*(h|k)eit(en)?/`
	`clean=/.*kait(en)?/ _i_ norm!=/.*(h|k)eit(en)?/`

-ung	
Initial sample	`norm=/.*ung(en)?/ _o_ clean _=_ dipl`
Basis sample	`norm=/.*ung(en)?/`
Graphematic variants	`clean=/.*ug(en)?/ _i_ norm!=/.*ung/`
	`clean=/.*nng(en)?/ _i_ norm!=/.*ung/`
	`clean=/.*umg(en)?/ _i_ norm!=/.*ung/`

I exported all samples from ANNIS using the CSV MultiTok Exporter with the parameters `metakeys=date, title, place, lang_type, lang_area`.

The CQL query used for the *-heit*/*-keit* sample from DECOW16B in Section 4.4.1 was `[lemma=".*(h|k)eit(en)?"]`.

	-er				*-heit/-keit*				*-ung*			
Step	Success	Fail	Outl.	Obs.	Success	Fail	Outl.	Obs.	Success	Fail	Outl.	Obs.
1	1600	400	157	1443	1719	281	378	1341	1732	268	226	1506
2	1581	419	335	**1246**	1670	330	270	1400	1636	364	147	1489
3	1764	236	296	1468	1650	350	218	1432	1548	452	98	1450
4	1729	271	269	1460	1560	440	192	1368	1487	513	77	1410
5	1726	274	266	1460	1494	506	181	1313	1471	529	70	1401
6	1766	234	294	1472	1452	548	172	**1280**	1442	558	58	**1384**
7	1753	247	269	1484	1448	552	128	1320	1465	535	49	1416
8	1743	257	254	1489	1403	597	107	1296	1438	562	47	1391
9	1726	274	235	1491	1441	559	103	1338	1454	546	56	1398
10	1688	312	210	1478	1496	504	99	1397	1533	467	56	1477
11	1693	307	192	1501	1594	406	104	1490	1673	327	58	1615
12	1677	323	185	1492	1701	299	110	1591	1839	161	35	1804
13	1706	294	182	1524	1800	200	99	1701	1962	38	60	1902
14	1688	312	173	1515	1871	129	104	1767	1990	10	55	1935
15	1693	307	161	1532	1926	74	94	1832	1992	8	60	1932
16	1698	302	159	1539	1944	56	71	1873	1994	6	62	1932
17	1763	237	132	1631	1979	21	61	1918	2000	0	63	1937
18	1804	196	120	1684	1994	6	65	1929	1998	2	51	1947
19	1911	89	127	1784	1998	2	64	1934	2000	0	39	1961
20	1990	10	95	1895	2000	0	47	1953	2000	0	73	1927

This table shows the number of successes, failures, outliers, and observed data points for the Monte Carlo analyses of S for each suffix from RIDGES. In boldface are the minimum numbers of observed data points for each suffix; the data for every other measurement point was cut down to this value, so that each measurement point has the same amount of data for the analysis.